A FIRST COURSE
IN
STATISTICS

BY

D. CARADOG JONES, M.A., F.S.S.
LECTURER IN SOCIAL STATISTICS AT LIVERPOOL UNIVERSITY
FORMERLY LECTURER IN MATHEMATIC
AT DURHAM UNIVERSITY

1386
LIBRARY
AGRICULTURAL COLL
20 11 25

LONDON
G. BELL AND SONS LTD.
1924

10,002—27-4-45.
AGRICULTURAL COLLEGE AND RESEARCH INSTITUTE
LIBRARY, COIMBATORE.

Acc No. Call. No.

Date Label

[This book should be returned on or
before the date last marked below.]

BELL'S MATHEMATICAL SERIES

ADVANCED SECTION

General Editor: WILLIAM P. MILNE, M.A., D.Sc.

Professor of Mathematics, University of Leeds

193

A FIRST COURSE IN STATISTICS

BELL'S MATHEMATICAL SERIES.

ADVANCED SECTION.

General Editor: WILLIAM P. MILNE, M.A., D.Sc.,
Professor of Mathematics, University of Leeds.

AN ELEMENTARY TREATISE ON DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS. By H. T. H. PIAGGIO, M.A., D.Sc., Professor of Mathematics, University College, Nottingham. Demy 8vo. Third Edition. 12s. net.

A FIRST COURSE IN NOMOGRAPHY. By S. BRODETSKY, M.A., B.Sc., Ph.D., Professor of Applied Mathematics, Leeds University. Demy 8vo. 10s. net.

AN INTRODUCTION TO THE STUDY OF VECTOR ANALYSIS. By C. E. WEATHERBURN, M.A., D.Sc., Professor of Mathematics, Canterbury University College, Christchurch, N.Z. Demy 8vo. 12s. net.

ADVANCED VECTOR ANALYSIS, with Application to Mathematical Physics. By the same Author. Demy 8vo. 15s. net.

A FIRST COURSE IN STATISTICS. By D. CARADOG JONES, M.A., F.S.S. Second Edition, revised. Demy 8vo. 15s. net. Also in 2 parts, Part I., 7s. 6d.; Part II., 9s.

THE ELEMENTS OF NON-EUCLIDEAN GEOMETRY. By D. M. Y. SOMMERVILLE, M.A., D.Sc., Professor of Mathematics, Victoria University College, Wellington, N.Z. Crown 8vo. 7s. 6d. net.

ANALYTICAL CONICS. By the same Author. Demy 8vo. 15s. net.

LONDON: G. BELL AND SONS, LTD.

First Published, . . 1921
Second Edition, revised, 1924

1386

Printed in Great Britain by T. and A. CONSTABLE LTD.
at the University Press, Edinburgh

PREFACE

FIFTY years ago a large section of the general public were not only uninterested in what we now call the social problem, but they scarcely gave a thought to the existence of such a problem. They felt vaguely perhaps, during periods of acute distress due to lack of employment, that all was not well and they thought the Government or possibly the big landowner was to blame, but only the more enlightened realized the complexity of the body politic and how fearfully and wonderfully it is made. To-day all this is changed, and comparatively few imagine that a single panacea—the prohibition of drink, the nationalization of land, or a levy on capital—will cure all evils.

The very fact that nearly the whole civilized world has given itself up for over four years to the destruction of life and the dragging down of the social fabric in all countries on so vast a scale has led to a surfeit and a reaction in which thoughtful men are eager to take part in proclaiming again a common brotherhood and in building a better world. Those who have always been interested in this kind of architecture welcome the change of spirit, but they also recognize the difficulty of the task undertaken and the need for no little mental effort to second the good-will, which is the first essential for success. To pull down no teacher is needed, but we must learn to build.

This leads one to the subject of the present book. The man who wishes his work to stand must make sure of its foundations. He cannot afford to rest satisfied, as too often the politician and social worker do, with wild and ill-informed generalizations where more exact knowledge is possible, and there are few human problems in the discussion of which some acquaintance with the proper treatment of statistics is not in the highest degree necessary.

Most people, however, are suspicious of figures. They imagine that quantitative considerations must of necessity deaden all feeling for the purely aesthetic or qualitative spirit which is the very life of the phenomena observed or measured. But this surely need not be the case. Kepler, when he succeeded in translating the motions of the planets into the language of number was not, we believe, the less but rather the more enamoured of the beauty and order with which the whole of creation is clothed.

A second reason for suspicion is that partisans of one school or another with more push than principle sometimes trade upon the general ignorance of statistics to 'prove' their own pet theories, while others no less enthusiastic lead the credulous public into the ditch, not with malice intent, but because they are really blind themselves to the right interpretation of the figures they so glibly quote.

Although a concern in social questions led the present writer in the first instance to study the theory of statistics, there is no reason why this bias should prevent the book being of service to those who wish to know something of its application in other directions, seeing that the general principles underlying the theory are the same in all cases, and illustrations have been taken from any field, biological, economic, medical, etc., just as they suited the immediate purpose in view.

The author makes no claim to any originality: he is no more than a student seeking to put together, with some kind of system and as he understands them, the simpler and more important ideas he has gathered from other sources. The matter is entirely the work of others, the manner only is his own, and he will be happy to receive criticism if thereby he may learn more. His chief qualification for writing is that he has had to worry through most of his difficulties alone, and consequently he knows where another student is likely to be in trouble better perhaps than the kind of writer who is so quick as to be able to see through things at a glance or, failing that, so fortunate as to be able to borrow immediate light from others.

The book is divided into two parts. Practically all the first part should be well within the understanding of the ordinary person.

Part II. is more mathematical, but an effort has been made throughout to explain results in such a way that the reader shall gain a general idea of the theory and be able to apply it without needing to master all the actual proofs. The whole is meant, not as an exhaustive treatise, but merely as a first course introducing the reader to more serious works, and, since real inspiration is to be found nowhere so surely as at the source, it is intended to encourage and fit him to pursue the subject further by consulting at least the most important original papers referred to in the text, only enough references being given to awaken curiosity. With the same intention a short chapter is inserted after the Appendix by way of suggesting a few of the sources of statistics likely to be of interest to the social student.

Some living writers, notably Professor Karl Pearson, have contributed so largely to the development and application of statistics that it is impossible to write upon the subject at all without incorporating large parts of their work, and the least one can do is gladly to record the benefit and pleasure one has received from them. The author's indebtedness to the two most important English text-books—Yule's *Theory of Statistics* and Bowley's *Elements of Statistics*—will be evident also to any one who knows these books, for they became so familiar through constant study that he fears he may have drawn upon them unconsciously even to the point of plagiarism in places.

Finally, he wishes specially to the acknowledge kindness of four friends—Mr. Peter Fraser, Lecturer in Mathematics at Bristol University, without whose encouragement in the early stages the work would never have been attempted ; Professor H. T. H. Piaggio, University College, Nottingham, and Mr. A. W. Young, sometime Lecturer at the Sir John Cass Technical Institute, London, whose criticisms and suggestions were most valuable ; and Professor W. P. Milne, of Leeds University, who, both as a practical teacher and as Editor of this series, ungrudgingly gave his help and advice.

D. C. J.

NOTE TO THE SECOND EDITION

THE kindly reception given to my book leads me to think that it might appeal to a wider circle of readers if they were not frightened by the mathematical appearance of certain pages in the second part. With this new issue it has been made possible to obtain Part I. and Part II. separately.

A selection of examples from London, B.Sc. (Econ.) papers has been included by kind permission of the Authority concerned. It is hoped that these may prove useful to students. An Index has also been added. Otherwise no changes of importance have been made in the text.

D. C. J.

September 1924.

CONTENTS

PART I

CHAP.		PAGE
I.	Introductory—Early Historical Beginnings: Logical Development	1
II.	Measurement, Variables, and Frequency Distribution	5
III.	Classification and Tabulation	14
IV.	Averages	22
V.	Averages (*continued*)—Applications of Weighted Mean	32
VI.	Dispersion or Variability	42
VII.	Frequency Distribution: Examples to illustrate Calculating and Plotting: Skewness	52
VIII.	Graphs—Correlation suggested by Graphical means	68
IX.	Graphs (*continued*)—Graphical Ideas as a Basis for Interpolation: Reasoning made Clear with the Help of Graphs or Curves	85
X.	Correlation	102
XI.	Correlation—Examples	115

PART II

XII.	Introduction to Probability and Sampling	132
XIII.	Sampling (*continued*)—Formulæ for Probable Errors	150
XIV.	Further Applications of Sampling Formulæ	165
XV.	Curve Fitting—Pearson's Generalized Probability Curve	178
XVI.	Curve Fitting (*continued*)—The Method of Moments for connecting Curve and Statistics	194
XVII.	Applications of Curve Fitting	206
XVIII.	The Normal Curve of Error	231
XIX.	Frequency Surface for Two Correlated Variables	249
APPENDIX		263
	Certain Current Sources of Social Statistics	279
	A Note on Tables to aid Calculation	284
MISCELLANEOUS EXAMPLES, Parts I and II		287
INDEX		299

PART I

CHAPTER I

INTRODUCTORY

Early Historical Beginnings. Statistics, more or less valuable, have been compiled in most civilized countries from very early times. One reason for doing this on a large scale has been to ascertain the man-power and material strength of the nation for military or fiscal purposes, and we read in the Old Testament of such censuses being taken in the case of the Jews, while among the Romans also it was a common practice.

In England, as economic terms began to be used and their meanings analysed, and especially during the period when the mercantile system prevailed, and the Government endeavoured so far as was practicable to direct industry into channels such that it would add most to the power of the realm, men tried frequently to base arguments for social and political reform upon the results of figures collected. A distinct advance had been made in the seventeenth century when mortality tables were drawn up and discussed by Sir William Petty and Halley, the famous astronomer, among others, and their labours prepared the way for a more scientific treatment of statistical methods, especially at the hands of one, Süssmilch, a Prussian clergyman, who published an important work in 1761.

It is almost true to say, however, that until the time of the great Belgian, Quetelet (1796-1874), no substantial theory of statistics existed. The justice of this claim will be recognized when we remark that it was he who really grasped the significance of one of the fundamental principles—sometimes spoken of as *the constancy of great numbers*—upon which the theory is based. A simple illustration will explain the nature of this important idea: Imagine 100,000 Englishmen, all of the same age and living under the same normal conditions—ruling out, that is, such abnormalities as are occasioned by wars, famines, pestilence, etc. Let us divide these men at random into ten groups, containing 10,000 each, and note the age of every man when he dies. Quetelet's principle lays

down that, although we cannot foretell how long any particular individual will live, the ages at death of the 10,000 added together, whichever group we consider, will be practically the same. Depending upon this fact insurance companies calculate the premiums they must charge, by a process of averaging mortality results recorded in the past, and so they are able to carry on business without serious risk of bankruptcy.

As a distinguished statistician once said, 'By the use of statistics we obtain from milliards of facts the grand average of the world.' But if the average resulting from our observations were subject to violent fluctuation as we passed from one set of facts to another cognate set there would be little satisfaction in finding it. It is the comparative constancy of the average, if the number of our observations is large enough, which makes it so important, as Quetelet observed, for although the idea was not altogether new he first realized how wide an application it had and how fruitful of practical results it might prove.

Quetelet was born in Ghent, and taught mathematics in the College there in his early youth. After graduating as Doctor of Science he became Professor of Mathematics in Brussels Athenæum when only twenty-three years old, and later he was made Director of the Brussels Observatory, in the foundation of which he had taken a leading part. In 1841 he was appointed President of the Central Commission of Statistics, where he was in a position to render valuable assistance to the Belgian Government by his advice on important social questions. He initiated the International Statistical Congress, which has served to bring together the leading statisticians of all countries, and the first meeting was held in 1853 at Brussels. His death occurred at the ripe age of seventy-eight.

Some idea of the extent of Quetelet's statistical researches may be gathered from the titles of his chief works: (1) *Sur l'homme et le développement de ses facultés, ou essai de physique sociale* (1835); (2) *Lettres . . . sur la théorie des probabilités appliquée aux sciences morales et politiques* (1846); (3) *Du système social et des lois qui le régissent* (1848); (4) *L'Anthropométrie, ou mesure des différentes facultés de l'homme* (1871).

In his writings he visualizes a man with qualities of average measurement, physical and mental (*l'homme moyen*), and shows how all other men, in respect of any particular organ or character, can be ranged about the mean or average man, just as in Physics a number of observations of the same thing are ranged about the mean of all the observations. Hence he concluded that the

methods of Probability, which are so effective in discussing errors of observation, could be used also in Statistics, and that deviations from the mean in both cases would be subject to the binomial law.

Hain in Vienna put some of Quetelet's ideas to good service in 1852, employing a superior method for the calculation of statistical variability. Knapp and Lexis in Germany, also following up Quetelet's principles, made an exhaustive investigation several years later of the statistics of mortality, and their work has been extended in many directions, and in our own time notably by Galton, Karl Pearson, and Edgeworth.

The name of Sir Francis Galton (1822-1911), to whose work as a pioneer the science of Statistics owes so much, is deserving of even greater honour than it has yet received. Founder of the School of Eugenics, Galton himself came of famous stock, being grandson of Erasmus Darwin and a cousin to Charles Darwin. He studied medicine in early youth, but after graduating at Cambridge his attention was turned to exploration, and the Royal Geographical Society awarded him a gold medal on the results of his investigations in South-West Africa. His first great work on heredity was not published till 1869, after he had already earned distinction in other directions, for he was elected a Fellow of the Royal Society in 1860. Alive with new ideas, marvellously patient and persistent in bringing them to the test of observation—qualities essential for real scientific research—he set himself to inquire into the laws governing the transmission of characteristics, physical and mental, from one generation to another. Large tracts of this ground have since been carefully explored and mapped out by the school of his great successor, Karl Pearson, who has originated formulæ for testing the extensive anthropometrical and biological data collected. Largely as a result of their work it is now widely recognized that 'the whole problem of evolution,' as Professor Pearson himself has well said, 'is a problem in vital statistics—a problem of longevity, of fertility, of health, and of disease, and it is as impossible for the evolutionist to proceed without statistics as it would be for the Registrar-General to discuss the national mortality without an enumeration of the population, a classification of deaths, and a knowledge of statistical theory.'

Logical Development. The best way to approach the study of any subject, if one had time, would be along the lines of its historical development, but these lines seem so often to diverge from the main theme, like branches from the parent stem of a tree, that

when one tries to describe them the general effect is apt to be somewhat confusing. It is therefore usually the custom to adopt a logical rather than a historical sequence, but it may assist the reader to see the connection between the two and the unity which embraces the whole if we now briefly trace the natural growth of the subject, suggesting the steps we might expect it logically to take. This we have tried to keep in view as nearly as possible in the succeeding chapters, except that the order may have been altered here and matter may have been omitted or inserted there as reason and the elementary nature of the work dictated :—

1. Owing to the difficulty which the mind experiences in grasping a large mass of figures, the necessity for an average arises to sum up shortly the character of the mass, and various kinds of averages are proposed.

2. An average proves insufficient alone to define the whole scheme of observations, and other constants are invented to measure their spread or dispersion about the average.

3. Considerations of space and the desire for some kind of system lead further to the formation of tables with the observations classified in ordered groups.

4. The formation of these tables suggests the possibility of a graphical representation of the numbers in the different groups to bring out the nature of their distribution.

5. The impossibility of dealing with a whole population results in the selection of samples, and the comparison of one sample with another introduces the subject of random errors.

6. The closer examination of this subject leads us into the domain of mathematical probability and discovers the probability curve, or normal curve of error, first formulated in connection with the study of errors of observation.

7. This same curve serves in the sequel to describe a certain important type of statistical distribution, in which each observation is determined by a multitude of so-called chance causes pulling this way and that, so that it is impossible to foretell what the resultant effect will be.

8. The failure of the normal curve to describe other common distributions, especially those which are unsymmetrical in character, leads to the development of skew varieties of curves which will fit them.

9. The extent of connection between one set of data and a possibly related set is a natural subject for inquiry giving rise to the theory of correlation.

CHAPTER II

MEASUREMENT, VARIABLES, AND FREQUENCY DISTRIBUTION

Measurement. There are two fundamental characteristics which pertain to nearly all measurement: it is (1) relative: it involves a comparison between one magnitude and another of the same kind, and (2) approximate: the comparison in practice cannot be made with absolute exactness.

A man's height, for example, is stated to be 5 ft. $8\frac{3}{4}$ in., but this would convey little to one who did not know how long a foot was and how long an inch was. The first step in the measurement is made by comparing the man's length with a certain constant length previously agreed upon as a standard or unit, namely, a 'foot'; he is placed to stand up against a scale which is divided up into feet, and the highest point of his head is seen to come somewhere between the 5 ft. line and the 6 ft. line: he is therefore longer than five of these units, set end to end, but not so long as six of them. To carry the measurement a stage further a smaller unit has to be introduced; each foot length of the scale is subdivided into twelve equal parts called inches, and the top of the man's head is found to come somewhere between the 5 ft. 8 in. line and the 5 ft. 9 in. line: he is therefore over 5 ft. 8 in., but not quite 5 ft. 9 in. in height. For the next stage in the measurement each inch of the scale has to be further subdivided into quarter-inches, and the top of the man's head is found to come somewhere between the 5 ft. 8 in. 3 qu. in. line and the 5 ft. 9 in. line; moreover it is nearer, let us suppose, to the former line than to the latter. In this case, then, we say that the man's height or length is 5 ft. $8\frac{3}{4}$ in., *measured to the nearest quarter inch.*

In measurement the decimal notation has very obvious advantages, because each unit is always divided into ten equal parts to get the next smaller unit. Thus a weight of 7 kilogr. 5 hectogr. 3 decagr. 8 gr. 4 decigr. 3 centigr. can be expressed at once in grammes, namely 7538·43 gr.; hence if we were measuring to the nearest decagramme, the result would be expressed as 754 decagr.; to the nearest decigramme, it would be 75384 decigr., etc.

Similarly, a length of 12 kilom. 7 metres 2 centim. can be written 12007·02 metres, or, in kilometres, 12·00702 kilom., or, to the nearest decametre, 1201 decam., and so on.

The mere act of counting things of a like kind is, in a sense, measurement of a primitive type, one thing being the unit, though the measurement may in many such cases be exact; for example, we may count the number of persons in a room exactly. Even in this type of case, however, the counting or measuring cannot always be done accurately, but the inaccuracy arises from lack of precision and uniformity in definition rather than from want of power in the measuring instrument itself: *e.g.* in determining the population of a city, inaccuracies may arise because of failure to define exactly the boundaries of the city, or the time at which the census is to be taken, or how to deal with the migration of the inhabitants from or into the city, and with births and deaths during the actual time of numbering.

Variables. *By a variable is meant any organ or character which is capable of variation or difference in size or kind.* The difference may be measurable as in the case of head-length, height, temperature, etc., or not directly measurable as in the case of colour, intelligence, occupation, etc. Further, the variation, when measurable, may be continuous, or it may take place only by integral steps, omitting intermediate values: population, for example, can never go up or down by less than one, but if temperature is to change from 60 degrees to 61 degrees it must pass continuously through every intermediate state of temperature between 60 degrees and 61 degrees.

In dealing with a measurable variable sometimes we are interested not so much in its actual value at a particular instant as in the change which has taken place in its value during some specified interval, but to gauge fairly the amount of this change it is necessary to measure it relative to the original value of the variable. For example, if we are told that the wages of a certain person have gone up during the year to the extent of 3d. an hour, we cannot say whether this is much or little to him until we know what his wages were originally. The addition would be relatively much less if he were a skilled patternmaker earning 1s. 6d. an hour than it would be if he were a chainmaker, earning only 6d. an hour.* This point can be met by stating, not simply the change in the value of the variable, but the ratio of the new value to the old. For instance, the patternmaker in the above instance has had his wages increased

[* Wages to-day are, of course, much higher—the above figures are only hypothetical.]

in the ratio of 1s. 9d. to 1s. 6d. It is important to notice that this form of measurement is quite independent of the particular units used; if we take 1d. as unit, the ratio=21/18=7/6, and if we take 1s. as unit, the ratio=$1\frac{3}{4}/1\frac{1}{2}$=7/6 just as before.

There are other ways of measuring this change in the value of a variable. One of the commonest is to express it as a percentage of the original value; thus the patternmaker's increase is at the rate of $\frac{3}{18}\times 100$, or $16\frac{2}{3}$ per cent., which is simply the ratio of increase in wage to previous wage multiplied by 100. The multiplier, 100, is quite an arbitrary factor, but it has obvious advantages: among others, it works well with the decimal notation and it often serves to put the result into a form which is greater than unity instead of leaving it as a fraction. Again, a man who gets a dividend of £25 on an investment of £500 receives interest at the rate of $\frac{25}{500}\times 100$, or 5 per cent.; in other words, this is the rate at which his capital accumulates if the interest is added to it instead of being spent.

Annual birth rates and death rates, on the other hand, are best expressed per thousand of the population, as estimated, say, at the middle of the year in question; *e.g.* the birth rate of the United Kingdom in 1911 was 24·4 per thousand, and the death rate was 14·8 per thousand, which is equivalent to 244 and 148 per 10,000 of the population respectively. If we could assume the birth and death rates to remain constant from year to year, and if we could afford to leave migration out of account, the population would be subject to exactly the same law of increase as capital accumulating at compound interest [see Appendix, Note 1], thus:—

1. If P be the original population, and if the annual net increase be at the rate of 25 per thousand, then

the population	in	1	year's time	$=P\times(1{\cdot}025)$
,,	,,	2	,,	$=P\times(1{\cdot}025)^2$
,	,,	3	,,	$=P\times(1{\cdot}025)^3$
,,	,,	n	,,	$=P\times(1{\cdot}025)^n$.

2. If £P be the original capital, and if the annual increase be at e rate of $2\frac{1}{2}$ per cent., then

the capital	in	1	year's time	$=P\times(1{\cdot}025)$
,,	,,	2	,,	$=P\times(1{\cdot}025)^2$
,,	,,	3	,,	$=P\times(1{\cdot}025)^3$
,,	,,	n	,,	$=P\times(1{\cdot}025)^n$.

Lest we may seem to have laboured to make plain what is really simple idea, it may be remarked that quite frequently confusion ises with regard to percentage even in reputable quarters. As an

illustration of the kind of mistake which, without thinking, is easily made, the following argument has been taken from a monthly circular sent out a little while ago to the members of the Boilermakers' Society by their Secretary: *Since July* 1914, *wages have risen* 15 *per cent., the cost of living has gone up* 45 *per cent., therefore the workers' real wages have fallen* 30 *per cent.* This same argument was quoted shortly after in one of the leading articles of *The Manchester Guardian* under the heading 'Prices and Wages,' and again in *The Labour Leader* tersely as truth 'In a Nutshell,' but in neither instance did it seem to have occurred to the writer that it was inaccurate. It may be worth while for the sake of clearness to show what the statement should have been :—

	Wages.	Cost of Living.	Ratio of Wages to Cost of Living.	Same Ratio multiplied by 100.
July 1914 . .	100	100	1	100
October 1916 .	115	145	$\frac{115}{145}$	79

Since $\frac{115}{145}\times 100$ is roughly 79, this calculation shows that 'real wages' had fallen only about 21 per cent. (100−79=21), and not 30 per cent. as stated, between the two dates.

Index Numbers. A very important case of variables changing with time appears in the discussion of changes in the value of money as measured by the movement of prices of commodities, introducing the notion of an index number. For example, supposing the wholesale price of beef was 6d. a lb. at one date, 8d. a lb. at another date, and 5½d. a lb. at a third date, the change might be exhibited as in the following table :—

	1st Date.	2nd Date.	3rd Date.
Price of beef .	6*d.*	8*d.*	5½*d.*
,, ,, .	100	133	92

Here 100, 133, and 92 are called index numbers, the price at the first date being taken as a standard and denoted by 100, while the prices at the other two dates are altered proportionally, so that

$$6 : 8 : 5\tfrac{1}{2} = 100 : 133 : 92.$$

Index numbers calculated on this principle have been published systematically for several years by Mr. A. Sauerbeck (in the *Journal*

of the Royal Statistical Society up to January 1913, and continued afterwards in *The Statist* under the supervision of Sir George Paish) and in *The Economist.*

In Sauerbeck's index numbers the average wholesale prices of forty-five commodities for the eleven years 1867-77 are taken as the standard, being denoted each by 100 as above, and the prices of the same commodities for any other year are then written as percentages of these standard prices. The commodities chosen are various—food of all kinds (cereals, meat, potatoes, rice, butter, sugar, coffee, tea), minerals (including coal), textiles, and sundries (including hides, leather, tallow, palm oil, olive oil, linseed, petroleum, soda, soda nitrate, indigo, timber). Articles of similar character are grouped together; naturally no class is exhaustive, but the selection is a fairly representative one. A sort of general average is then formed by combining all the results, and the movement of this average is taken to measure changes in the value of money. An example will make clear the way in which an index number for each group and the general average are obtained.

The index number for each separate commodity may be first calculated thus:—

PRICE OF ENGLISH WHEAT.

Years.	Price per Quarter.	Index Number.
	s. d.	
1867-77 . .	54 6	100
1912 . . .	34 9	64

Now forming similar index numbers for each of the eight vegetable and cereal foods and combining them together, we have:—

INDEX NUMBERS FOR VEGETABLE AND CEREAL FOODS.

Years.	English Wheat.	American Wheat.	Flour.	Barley.	Oats.	Maize.	Potatoes.	Rice.	Eight Commodities.	Average.
1867-77 .	100	100	100	100	100	100	100	100	800	100
1912. .	64	68	70	79	83	85	74	101	624	78

The figures in the last column but one are obtained by simply adding the figures in the eight previous columns, and, dividing these

results by eight, we get the average index number for the group in 1912 as a percentage of that in the standard years 1867-77.

Treating all the other commodities in the same way we ultimately get index numbers for all the different groups and for all commodities combined as follows :—

INDEX NUMBERS FOR DIFFERENT GROUPS AND FOR ALL COMMODITIES.

No. of Commodities	8	7	4	19	7	8	11	45
Years.	Vegetables and Cereals.	Animal Food.	Sugar, Coffee, Tea.	All Food.	Minerals.	Textiles.	Sundries.	All Commodities.
1867-77	100	100	100	100	100	100	100	100
1912	78	96	62	81	110	76	82	85

The index number for 'All Food' is obtained by summing the nineteen index numbers for the separate commodities which are included in this class and dividing the result by 19. Similarly the general index number for all commodities is obtained, not by adding the numbers for the different groups and dividing by the number of groups, but by adding the forty-five index numbers of all the separate commodities and dividing the result by 45.

In *The Economist* the average prices of twenty-two commodities for the years 1901-5 are taken as the standard, being denoted each by 100, and the prices of the same commodities for any other year are then written as percentages of these standard prices ; the sum of these percentages is taken as the index number, and it is a simple matter to divide by 22 if we wish to get the *average* percentage change. The following table explains the method of calculation :—

INDEX NUMBERS REPRESENTING PRICES OF COMMODITIES

Date.	Cereals and Meat.	Other Foods.	Textiles.	Minerals.	Miscellaneous.	Total.	Index No. / 22.
1901-5 . .	500	300	500	400	500	2200	100
End of Dec. 1916	1294	553	1124·5	824·5	1112	4908	223

In this table five commodities are included under the head of 'Cereals and Meat,' three under 'Other Foods,' and so on. The numbers in the last column are obtained by dividing those in the previous column by 22.

It is clear that what is at bottom the same principle may be applied in any case of a variable changing with time when we wish to measure the extent of the change, so that the use of index numbers is not confined to the problem of prices. We shall return again to discuss one or two further points in connection with the same subject in the Chapter on 'Averages.'

Frequency Distribution. So far we have been thinking more particularly of the change which an individual variable, or a collection of such variables, may undergo in the course of time, or the difference between two values which the same variable may have at two different instants of time, and how to measure it. Now the science of Statistics is based upon the study of the crowd rather than of the individual, although observations on individuals have to be made before they can be combined together to produce the crowd, just as individual income-tax schedules have to be completed and combined before the balance-sheet of the State can be drawn up. As we pass from one individual to another there may be great differences in the organ or character observed—hence the word variable already introduced—but in the mass these differences are merged together and lose their individual importance: it is rather their resultant effect we seek to measure. In order therefore to discover this effect it is necessary to make a collection of individual observations and to analyse the results. Now if our ultimate conclusions are to be safe the number of observations must be considerable, and in order to be able to cope with them and reduce them to some sort of system the first step in the analysis consists in arranging them in different classes according to the value of the variable under consideration.

It is to be noted that now we are dealing with changes in the value of a variable as we pass *from one individual to another at the same period of time* and under the same general conditions, and not with the change in a variable *in the same individual occurring with the lapse of time.* We wish, for example, to draw a distinction between (1) the change in wages as we pass from one man to another at the same time in the same trade, and (2) the change in wages of the same man, or class of men, in the same trade occurring in a given period of time; in the first case we want to find the amount

of diversity within the trade at some stated time, and in the second our object is to discover whether an improvement has taken place in the wages of a particular individual or a particular trade with the passage of time.

In picturing variation of the first type the conception arises of a *frequency distribution* where the observations are distributed in ordered groups, with a number corresponding to each showing how many, or how frequent, are the individuals possessing the type of variable or character which defines that group. More generally, if a series of measurements or observations of a variable y are made corresponding to a selected series of another variable x we get a distribution, which becomes a frequency distribution when y represents the frequency of events happening in a particular way, or of individuals corresponding to a particular value of some common variable or character, represented by x. Thus (1) the boys in a school might be grouped according to their intelligence: so many, dull; so many, of ordinary intelligence; and so many, bright or above the ordinary. Again (2) in an inquiry into the housing of the people in any town or district it would be necessary to draw up a table showing the number or frequency of existing tenements with one room, the frequency of tenements with two rooms, the frequency of tenements with three rooms, and so on. Once more (3) a zoologist, wishing to discover whether crabs of a certain species caught in one locality differ in any remarkable way from members of the same species caught in another locality, might start by making measurements of the length of carapace or upper shell for crabs of like sex in the two places and then proceed to form frequency tables for each, setting out the frequency of crabs for which the carapace length lies, say, between 5 and 6 millimetres, the frequency with length between 6 and 7 millimetres, the frequency with length between 7 and 8 millimetres, and so on. He would then have in these tables some basis for comparing the specimens caught in the two localities.

The three illustrations just used give three different types of distribution corresponding to the three types of variable to which attention has been drawn before. In the first, where the variable or character observed is not measurable, doubt will sometimes arise as to the appropriate class in which individuals should be placed who seem to be on the border line between dulness and mediocrity or between mediocrity and brilliance, so that accurate classification will greatly depend upon what is called the 'personal equation' of the observer. The second illustration corresponds

to the case where the variable changes not continuously but by unit stages; the choice of classes in such a case depends little upon the observer unless the unit is very small compared to the total range of variability; for example, a tenement might either definitely have two rooms or it might have three rooms, but it clearly could not be put down as having $2\frac{1}{4}$ rooms or $2\frac{5}{6}$ rooms: in other words, the only natural classification is so many tenements with two rooms, so many with three rooms, so many with four rooms, and so on, though here too some confusion might arise through failure to define clearly what is 'a room.' In the third type, where we can conceive of the continuous variation of the character under observation, there would be nothing surprising in the appearance of any value of the variable between the lowest and highest values observed; the choice of suitable limits for the several groups becomes therefore in this case rather a delicate matter which requires careful judgment.

We shall begin the next chapter with some general remarks upon the subject of classification and tabulation.

CHAPTER III

CLASSIFICATION AND TABULATION

No part of Statistics is of more importance than that which deals with classification and tabulation, and it is the one part for which no very precise rules can be given. A neat arrangement of ideas in the mind, capacity to express them clearly, and patience are indispensable, but experience alone will convince one of the extreme care which must be exercised if blunders are to be avoided and time is to be saved in the long run. This has to be emphasized because most people, until they have tried and failed, imagine that to arrange things in classes and in tables is a straightforward proceeding involving no great thought or trouble.

Abundant matter of a statistical character is published periodically in Blue-books, Government Reports, Reports of Local Authorities, Directors of Education, Medical Officers of Health, Chief Constables, Employers' Associations, Trade Unions, Co-operative Societies, etc., but it needs a trained intelligence as a rule to assimilate it and turn it to further advantage. The larger the scale upon which any inquiry is made, the more valuable should the results be, granted that equal accuracy is possible on the large as on the small scale, but it is fairly clear that mistakes of various kinds have also much more chance of creeping into a large work than into a small one. To appreciate the various and numerous possibilities of error when the scope is wide it is enough to read the introductions to the Registrar-General's Reports on the Census from decade to decade ; this should also impress the student with the care that is necessary if he proposes to use such material for the investigation of some other problem. It may seem a comparatively simple task to abstract two sets of figures from a Census Report, to establish a one-to-one correspondence between them, and to make deductions therefrom, but such figures when taken from their context will sometimes lead to absolutely unsafe, if not false, conclusions. The exact meaning and limitations of any data can only be properly appreciated by one who has been closely in touch with the persons who have collected them, and it is therefore important, before

attempting to re-classify or re-tabulate any old statistics for a new purpose, to read very carefully through the notes made by the original compilers.

Perhaps the best advice that can be given to any one in this connection is that he should embark upon some small inquiry which will necessitate the collection of statistics for himself; the final result of his efforts may seem disappointing, but the experience he will gain will be invaluable. Ideas for such an inquiry will occur to him if he reads through some authoritative work on social questions, *e.g.* Beveridge's *Unemployment*, the decennial *Census Reports*, or *The Minority Report on the Poor Law* (1905). But he must read with an open and critical mind, questioning particularly the foundation for all statements as to cause and effect which may be made. A few simple hints may be useful as to method of procedure.

When he thinks he has discovered some subject of interest which would appear to deserve examination, it will be well to put it down on paper in order to get it clearly defined, because a precise written statement is likely to carry one further than a shadowy idea somewhere at the back of the mind which is hardly formulated at all. When the actual collection of statistics is begun it will almost certainly be found that it is impossible to solve the original problem contemplated; but that need not prevent further progress—what is important is that the limitations should be exactly realized, and this will be impossible unless the original problem is clearly presented side by side with the nearest solution obtainable.

The problem stated, the next thing is to set down categorically a number of questions, the answers to which are to be the raw material for the solution of the given problem. For the answers let us assume the inquirer is dependent upon the goodwill of others, either employers, or trade union secretaries, or public officials. The questions in that case must be clearly, concisely, and courteously phrased, and must not be capable of more than one interpretation. In number they should be few and in character not inquisitorial; moreover, the replies should be obtainable without any great labour on the part of the persons approached. Here again it will be found that the questions first set down are not all satisfactory: one will be too vague; another, though clear enough, may involve a considerable search through a mass of other matter before it can be properly answered; while to another it might be impossible to give an exact reply in any case. Revision and amendment may there-

fore be necessary in the light of the first replies received, and the inquirer will begin to see at this stage how far the solution to his original problem is really possible.

When the bulk of the returns have come in they should be critically examined one by one. A number will, for one reason or another, be worthless, and they must be discarded; as for the remainder, if the questions were well chosen, the answers should not be difficult to interpret and classify; the most successful questions are those to which a simple 'yes' or 'no' in reply gives all the information required; numerical answers are less easy to deal with, especially if there is the least chance of misunderstanding on either side as there often is, for example, in the case of observations which are on the border line between two classes.

Tables should then be drawn up and the headings to the different columns of the tables should state concisely and exactly what the figures below represent. So far as possible any one should be able readily to grasp their general meaning without being obliged to wade through a page or two of written explanation; if any heading cannot be clearly expressed in a few words it may be helped out by a further note at the bottom of the page, but too many such notes are to be avoided.

Finally, a summary should be made of the various conclusions suggested by a study of the tables. Some of the points raised in the course of the inquiry will perhaps be only incidental to the main problem under discussion, but may still deserve a passing reference. It will also be of advantage to follow up the summary by any recommendations which can be fairly based on the conclusions obtained, when the problem is such that recommendations are expedient, and, if ultimately the whole is of sufficient value to be printed, emphasis can be introduced where necessary by suitable variations in type.

For this part of the work considerable judgment is necessary which can only be acquired by long training—a faculty to pick out the real from the false and an eye to distinguish the important from the trivial. A sense of numerical proportion too is desirable incidentally; one of our leading exponents on finance in a book dealing with the meaning of money uses a very interesting illustration which is perhaps worth quoting here to show how even an acute mind may on occasion prove itself curiously lacking in such a sense. He is seeking to show how the credit system of the country is built upon a foundation composed of a little gold and a lot of paper; for this purpose he amalgamates together the balance-sheets of half

a dozen big banks, and proves that their liabilities on current and deposit account amounted at a certain date prior to 1914 to 249 million pounds, while the cash in hand and at the Bank of England was 43 millions. Of the 43 millions he estimates that roughly 20 millions would be cash in the Bank of England, and further that about two-thirds of this 20 millions would be represented really by securities and not by gold. Hence he concludes that *to support this vast erection of credit there would only be* £6,666,666 *of actual gold.* Thus after talking throughout in millions the author closes by giving his result true apparently to a pound!

Much may be learnt as to methods of classification and the drawing up of tables by a careful study of those which appear in various official reports, and a few such tables are reproduced in the pages which follow.

TABLE (1). CONDITION AS TO CLEANLINESS OF SCHOOL CHILDREN IN SURREY.

Cleanliness.	5 years, 1908-12. 79,070 children inspected.
Above the average .	15·4 per cent.
Average . .	76·5 „
Below average .	7·6 „
Much below average	0·5 „

TABLE (2). CONDITION AS TO INFECTIOUS DISEASES OF SCHOOL CHILDREN AT DIFFERENT AGES IN SURREY (1913).

Age Groups inspected	5-6	8-9	13-14	Total at All Ages.
Numbers inspected	5,191	5,151	4,962	15,304
Proportion who before inspection had suffered from—	per cent.	per cent.	per cent.	per cent.
Diphtheria . . .	1·3	3·5	5·4	3·4
Scarlet fever . .	2·7	7·2	10·9	6·9
Measles . . .	55·3	79·3	84·6	72·9
Whooping cough . .	41·8	56·4	54·3	50·9
German measles . .	2·9	5·1	7·5	5·1
Chicken pox . .	26·1	40·1	38·6	34·9
Mumps . . .	10·6	22·0	29·8	20·7
No infectious diseases .	18·9	6·1	4·7	10·0
No definite information	3·3	2·2	0·9	2·2

TABLE (3). HEIGHT OF SCHOOL CHILDREN ACCORDING TO DISTRICT, AGE, AND SEX (1913).

Age Groups.	Boys.				Girls.			
	Nos. measured.	Average Height in inches.	Average Height in cms.		Nos. measured.	Average Height in inches.	Average Height in cms.	
			Surrey.	England and Wales.			Surrey.	England and Wales.
5-6	2724	41·4	105·2	103·4	2467	41·3	104·9	102·6
8-9	2578	47·8	121·4	120·4	2573	47·5	120·7	119·4
13-14	2529	57·0	144·8	142·4	2433	57·9	147·1	144·2

The first four are taken from the *Annual Report of the School Medical Officer for the County of Surrey*, 1913. The first is an example of single tabulation showing the distribution according to cleanliness of children inspected in the elementary schools. The second is an example of double tabulation, showing the distribution according to age of school children who at some period before the date of inspection had suffered from certain infectious diseases. The third is an example of quadruple tabulation, showing the distribution of school children according to height, district, sex, and age. Thus in the first case we have one factor brought into relief, viz. cleanliness; in the second case we have two factors, age and disease; in the third case we have four factors, height, district, sex, and age.

When we have two or more factors tabulated together as in cases (2) and (3), we may be sometimes led to discover a connection of some kind, possibly causal, between them, and the search for such a connection, or *correlation* as it is called, represents one very useful purpose to which tabulation may be put. Table (4) is an illustration of this. It is the result of certain measurements carried out in order to discover the effect of employment out of school hours upon the physical condition of boys. The particular factor examined as the possible cause of evil in this connection is lack of sleep, and the figures given certainly seem to warrant a closer examination into the matter.

TABLE (4). PHYSICAL CONDITION OF CERTAIN BOYS ACCORDING TO HOURS OF SLEEP OBTAINED.

No. of Hours Sleep obtained.	No. of Boys examined.	Average Height in inches.	Average Weight in lbs.	Nutrition.		
				Percentage above average.	Percentage average.	Percentage below average.
7 to 8	14	54·5	71·3	7·1	35·8	57·1
8 to 9	80	55·4	73·9	10·1	65·9	24·0
9 to 10	296	56·4	79·3	15·3	64·5	20·2
10 to 11	280	57·9	83·2	22·8	66·5	10·7
11 to 12	50	59.0	87·0	22·0	68·0	10·0

Tables (5) and (6) are two illustrations of neat tables, containing a large amount of information in a small space, set out in such a form that the eye can easily take it in—and that is the main purpose of tabulation. These examples are selected from the *Sixteenth Abstract of Labour Statistics of the United Kingdom*, Cd. 7131.

In Table (6) note the classification of age groups: it is not '5 to 10 years,' '10 to 15 years,' and so on, but '5 and *under* 10 years,' '10 and *under* 15 years,' and so on. This removes difficulties at the border lines between two classes; the difficulties are not completely removed, however, unless there is some understanding as to what shall constitute *under* any particular age. Shall it be six months under, or one day under, or one hour under? This sort of ambiguity has more importance in some cases than in others. Suppose, for example, we were classifying men according to their height: a group of the type '60 inches and under 62 inches,' assuming that measurements were made to the nearest half-inch, would really include all men who were '59¾ inches and under 61¾ inches'; because one who measured anything from 59¾ in. to 60¼ in., being nearer to 60 in. than to 59½ in. measuring to the nearest half-inch, would be registered as 60 in. in height, while one who measured anything from 61¾ in. to 62¼ in., being nearer to 62 in. than to 61½ in., would be registered as 62 in. in height.

Another point to be noted is that in general people making returns seem to have a psychological weakness for round figures, so that a man in the neighbourhood of 40 years of age, for example, is apt to record himself as actually 40 although he may really

Table (5). Classification of Overcrowded Tenements—* England and Wales (1911).

Tenements with	Urban Districts. No. of Overcrowded Tenements.	Urban Districts. Occupants thereof. No.	Urban Districts. Occupants thereof. Percentage of total population.	Rural Districts. No. of Overcrowded Tenements.	Rural Districts. Occupants thereof. No.	Rural Districts. Occupants thereof. Percentage of total population.	Total. No. of Overcrowded Tenements.	Total. Occupants thereof. No.	Total. Occupants thereof. Percentage of total population.
1 room .	56,290	206,022	0·7	1,545	5,748	0·1	57,835	211,770	0·6
2 rooms .	119,695	712,613	2·5	15,397	91,458	1·2	135,092	804,071	2·2
3 rooms .	107,892	847,937	3·0	22,380	175,988	2·2	130,272	1,023,925	2·8
4 rooms .	64,470	624,747	2·2	17,341	167,969	2·1	81,811	792,716	2·2
5 or more rooms .	21,200	251,405	0·9	4,700	55,585	0·7	25,900	306,990	0·8

Table (6). Population grouped according to Age— England and Wales (1911).

MALES.

Age Groups.	Urban Districts. Number.	Urban Districts. Percentage.		Rural Districts. Number.	Rural Districts. Percentage.		All Districts. Number.	All Districts. Percentage.	
nder 5 years	1,517,432	11·3	41·2	418,681	10·6	41·2	1,936,113	11·1	41·2
5 and under 10 years	1,431,900	10·6		415,395	10·5		1,847,295	10·6	
,, 15 ,,	1,341,586	9·9		406,045	10·3		1,747,631	10·0	
,, 20 ,,	1,267,500	9·4		387,395	9·8		1,654,895	9·5	
,, 30 ,,	2,332,135	17·3	44·4	626,300	15·9	40·9	2,958,435	17·0	43·6
,, 40 ,,	2,094,934	15·5		542,370	13·7		2,637,304	15·1	
,, 50 ,,	1,556,818	11·6		444,360	11·3		2,001,178	11·5	
,, 60 ,,	1,042,868	7·7	14·4	333,368	8·4	17·9	1,376,236	7·9	15·2
,, 70 ,,	612,741	4·5		230,306	5·8		843,047	4·8	
and upwards	296,246	2·2		147,228	3·7		443,474	2·5	
Total .	13,494,160	100·0		3,951,448	100·0		17,445,608	100·0	

* For the purpose of the Census Reports 'ordinary tenements which have more than two occupants per room, bedrooms and sitting-rooms included,' are considered overcrowded.]

be 39 or 41 years old. To diminish the error arising from this fact it is usual, when not otherwise inconvenient, to fix the centres of the class-intervals at round figures : *e.g.* to take ' 15 and under 25 years,' ' 25 and under 35 years,' etc., in preference to ' 20 and under 30 years,' ' 30 and under 40 years,' etc. Where there is any known bias in the data, as, for instance, in the familiar case of certain women who consistently register themselves as younger than they really are, a correction can be made in the final figures.

In any frequency distribution where we wish to group a number of observations according to the magnitude of some common variable, as in Table (6) a number of males grouped according to age, the question arises—' How many groups should there be ? ' With this question is involved also the size of the corresponding class-interval, and this should be so large that, with possible exceptions at either extremity of the table, there are a fair proportion of observations to each class or group ; and, contrariwise, it should be so small that all the observations in any one group may be treated practically as if they were located at the centre of the group so far as the variable in question is concerned, *e.g.* it should be possible to treat males recorded in class ' 50 and under 60 years,' where the interval is 10 years, as if they were all of age 55 years. It will be found in general that a number of groups somewhere in the neighbourhood of 20 is the most satisfactory, granted that the number of observations is reasonably large, although in some cases it is impossible to split up the unit of class-interval, and we are obliged to be satisfied with a smaller number of groups on this account : Table (5) is a case in point where we are tied down to one room as the class-interval. In Table (6) the class-interval varies, being only 5 years at first, and afterwards 10 years, but as a rule the labour of calculation of the different statistical constants we require is considerably simplified if it is possible to keep the size of the class-interval the same for each group.

CHAPTER IV

AVERAGES

Common Average or Arithmetic Mean. Let us consider one of the commonest meanings of the term *average*. If a train travels a distance of 180 miles in 3 hours we say that it has been moving at 60 miles an hour. By this we do not mean that its speed is always 60 m/h, never more, never less, but that *if* it had moved always at that uniform speed it would have accomplished its journey in exactly the same time. As a matter of fact, during some instants it may have been moving at a much slower rate than 60 m/h, but, if so, it must have made up for this slackness by travelling at a much faster rate than 60 m/h during other instants, so that on the whole a balance was effected, and, as we say, the speed averaged out at 60 m/h.

Again, suppose the wages of three men are: A, 27s. a week; B, 18s. a week; C, 30s. a week. We should say that the average wage of the three was equivalent to

$$\tfrac{1}{3}(27+18+30)\text{s.}=25\text{s. a week.}$$

In other words, if A, B, and C were all under the same employer, and if, instead of paying them different amounts, he wanted to pay them all equally, he would have to give each man 25s. a week, assuming that his total wages bill was to remain unaltered. This method of measurement gives what is known as the *arithmetic mean*, or, more simply, *the mean.*

Once more, in discussing the state of the labour market as regards different trades, when we wish to compare one with another, it is not the actual numbers unemployed in each trade that are quoted, but these numbers expressed as percentages of the total numbers employable in each trade.

In each of these three cases we reduce our observations or measurements to a sort of common denominator, so that they may be mentally compared or contrasted more readily with other observations of a similar character. Thus we have in mind a certain mean

train speed per hour, or mean wage per week, or mean percentage out of work, as the case may be.

An average then in general we may regard as one of a class of statistical constants (others of which we are to meet later) which concisely label a set of observations or measurements pertaining to a common family. It is designed to describe the family type more nearly than is possible by observing any chance member, and in value it should therefore come somewhere near the middle of the family group, so that if the individual members of the family chance to be equal each to each in respect to the organ or character observed it should have the same value as they have. This constitutes a test for the validity of any formula giving the average of a set of observations: *e.g.* we might, if we wish, define the average of three numbers, p, q, r to be, not $\frac{1}{3}(p+q+r)$ but

$$\sqrt[3]{\tfrac{1}{3}(p^3+q^3+r^3)},$$

for (1) this formula, too, can be shown to give a number intermediate in value between the greatest and least of the numbers p, q, r; also (2) if we put $p=q=r=k$ (say), the formula reduces to

$$\sqrt[3]{\tfrac{1}{3}(k^3+k^3+k^3)}=\sqrt[3]{k^3}=k.$$

Clearly the range of choice for the definition of an average is infinite, though only a few definitions give averages which have proved their utility and come into general use. Of these the most important is the common mean already introduced, with its extension, the weighted mean, but at least two others deserve special consideration, the median and the mode.

Median. In any observed distribution if all the individuals can be arranged in order of magnitude of the character or organ observed, which may be conveniently done when they are not very numerous, the median organ or character will be that pertaining to the individual half-way along the series, so that there are in general an equal number of individuals above and below the median. For instance, if seven boys of different heights be placed to stand in a row, the tallest first, the next tallest next, and so on, the median height is the height of the fourth boy from either end. If there are an even number of boys, say eight, it would be natural to take as median the height midway between that of the fourth and that of the fifth boy.

When the items are numerous they are frequently grouped into classes, as we have seen, such that all in the same class are reckoned

to have some value lying between the extreme limits of that class. We should then, as before, halve the total number of observations to fix the particular individual which defines the median organ or character. This would enable us to pick out the group in which the median lies, and on reference to the original record of observations, assuming it was at hand, it would be a simple matter to identify the median.

If the original record be not available, however, it will be necessary to proceed to get the best value we can for the median in some other way. Consider, for example, Table (7), showing the distribution of marks obtained by 514 candidates in a certain examination. We begin by rearranging the data in the manner shown below Table (7). Now in accordance with the definition the median in marks should, strictly speaking, be midway between the marks assigned to the 257th candidate and the marks assigned to the 258th candidate : in fact, the marks corresponding to candidate number 257·5, if it were possible for such a candidate to exist. But we are ignorant so far as Table (7) goes of the marks gained by either the 257th or the 258th candidate, though it is possible, by the simple proportional process known as 'interpolation,' to calculate approximately the marks we require. We think of all the candidates as forming an ordered sequence, ranged one after the other according to their marks just like the boys of different heights, and the table shows that in this mental picture

the 231st candidate gets approximately 30 marks, while
,, 318th ,, ,, ,, 35 ,,

Hence candidate number 257·5, if one existed, ought to get a number of marks somewhere between 30 and 35. But, in this neighbourhood of the sequence,

a difference of (318-231) candidates corresponds to a difference of 5 marks, therefore

a difference of (257·5-231) candidates corresponds to a difference of $(\frac{5}{87}\times 26{\cdot}5)$ marks.

Thus the marks obtained by candidate number 257·5 are approximately

$$=30+\tfrac{5}{87}\times 26{\cdot}5$$
$$=31{\cdot}523,$$

and this may be taken as the median.

On examining the actual marks-sheet it was found that 252 candidates obtained 31 marks or less, and 273 candidates obtained

32 marks or less, so that the real median was 32, because this was the number of marks gained by both the 257th and the 258th candidates. The number 31·523 found above, however, would be a good approximation to take for the median when all the information at our disposal was that shown in Table (7).

TABLE (7). MARKS OBTAINED BY 514 CANDIDATES IN A CERTAIN EXAMINATION.

Marks Obtained.	No. of Candidates.	Marks Obtained.	No. of Candidates.
1 to 5	5	36 to 40	79
6 to 10	9	41 to 45	50
11 to 15	28	46 to 50	37
16 to 20	49	51 to 55	21
21 to 25	58	56 to 60	6
26 to 30	82	61 to 65	3
31 to 35	87		
		Total .	514

The table is to be read as follows :—

5 candidates obtained 1, 2, 3, 4, or 5 marks,
9 „ „ 6, 7, 8, 9, or 10 „ and so on.

By straightforward addition it can evidently be rearranged so as to read thus :—

5 candidates obtained not more than 5 marks.
14 „ „ „ „ 10 „
42 „ „ „ „ 15 „
91 „ „ „ „ 20 „
149 „ „ „ „ 25 „
231 „ „ „ „ 30 „
318 „ „ „ „ 35 „
397 „ „ „ „ 40 „
447 „ „ „ „ 45 „
484 „ „ „ „ 50 „
505 „ „ „ „ 55 „
511 „ „ „ „ 60 „
514 „ „ „ „ 65 „

It will be noted that in calculating the median no use is made of the marks of any of the candidates except those in the two groups in the immediate neighbourhood of the median, and it is one of the great advantages of this average that it can be found when an exact knowledge of the characters of the more extreme individuals in the series is not in our possession, and even when their measurement is impossible: it is enough if they can be roughly located. The arithmetic mean on the other hand is often unduly influenced by abnormal individuals which are not really typical of the population in which they appear.

Mode. If we measure or observe some organ or character for each individual in a given population, the mode, as its name suggests, is simply the organ or character of most fashionable or most frequent size. A large draper, for example, will have collars of several different shapes and sizes in his shop, but the fashionable shape and the predominant size correspond to the mode: it is the mode that sells most readily, and the intelligent draper will always have it in stock. Again, in Table (2), the disease mode or fashionable disease among certain school children inspected in Surrey in 1913 was measles, for a greater percentage of children had suffered from measles than from any other of the diseases recorded.

Now when the variable in which we are interested is 'discrete,' that is, when it changes by unit steps, leading to classes like 'tenements with 1 room,' 'tenements with 2 rooms,' 'tenements with 3 rooms,' and so on, it is an easy matter to pick out the class of greatest frequency: thus, in Table (5) there are more overcrowded tenements with 2 rooms than with any other number of rooms in the urban districts, so that 2 is the mode so far as this character (number of rooms) is concerned, whereas in the rural districts 3 is the mode, for there are more overcrowded tenements with 3 rooms than with any other number. There may be ambiguity, however, in determining the mode in this way for a grouped frequency distribution when we are dealing with an organ or character subject to 'continuous variation.' To cover such cases the modal value has been defined as that value for which the frequency per unit variation of the organ or character is a maximum. The precise significance of this wording will only be appreciated after discussing frequency curves: at present it must suffice to give a practical illustration of how the ambiguity arises and calls for some more refined treatment.

For this purpose turn again to the examination marks in Table (7),

from which it appears that the mode, if it is to be the marks obtained by the greatest number of candidates, should lie in the group (31 to 35), since there are 87 candidates with marks between these limits, and this number exceeds that in any other group. But how are we to decide the exact point in the interval (31 to 35) which is to correspond to the mode? Shall it be 33? We might say 'yes' if the distribution were perfectly symmetrical on either side of the (31 to 35) group, but if we examine the neighbouring groups we see that the balance leans rather more heavily to the (26 to 30) group with a frequency of 82 than to the (36 to 40) group with a frequency of 79, and we might allow for this by interpolating in some way—ignoring, of course, any errors which may occur in the frequencies themselves owing to the observations being generally limited in number. But the pull in the direction of lower marks becomes still more pronounced to our minds when we contrast also the frequencies in the next groups on either side, namely 58 and 50. So we might go on until the influence of the whole field of observations comes into action.

Now it so happened that in this particular case the original marks-sheet was to be seen, and a regrouping of the candidates as in Table (8) makes it clear that the value found in this way for the mode may be artificially displaced sometimes to a serious extent by the particular method of grouping adopted. Thus, according to this new arrangement, the mode would seem to lie in the interval (28 to 32), the mid-value of which differs materially from 33, the mid-value of the previous maximum frequency group.

TABLE (8). MARKS OBTAINED BY 514 CANDIDATES IN A CERTAIN EXAMINATION (ALTERNATIVE GROUPING).

Marks Obtained.	No. of Candidates.	Marks Obtained.	No. of Candidates.
3 to 7	10	38 to 42	73
8 to 12	17	43 to 47	45
13 to 17	35	48 to 52	31
18 to 22	56	53 to 57	12
23 to 27	47	58 to 62	3
28 to 32	108	63 to 67	3
33 to 37	74		
		Total .	514

[It should be observed that while an alteration of the grouping may also affect the median, it does not affect it nearly to the same extent: *e.g.* the median determined from Table (8) is 31·3, which differs little from 31·5 the value obtained by the first grouping.]

If, again, we combine the results of our two groupings to find the mode we might be tempted to conclude that it lies somewhere between the limits 31 and 32, but on examining the original records it was discovered that the real mode was 28. The frequency distribution of candidates in this neighbourhood was in fact very interesting; it ran as follows:—

Number of	candidates who	obtained	25	marks	=14
,,	,,	,,	26	,,	=10
,,	,,	,,	27	,,	= 6
,,	,,	,,	28	,,	=33
,,	,,	,,	29	,,	=17
,,	,,	,,	30	,,	=16

The explanation of this peculiar distribution seemed to be that 28 marks were required for a candidate to pass, and apparently as many candidates as possible were pushed over the pass line: if, on the first marking, a candidate was found to want only one mark to pass, the examiner presumably looked through his paper again and did his best to find an answer which by kindly treatment might be granted an extra mark. The effect of this leniency was ultimately to leave only 6 candidates in the division immediately below the pass line, and to swell the number immediately above to 33, which thus made 28 easily the 'most fashionable' mark of any, the next largest group of candidates being only 21. It will be observed that even a candidate who wanted 2 marks to pass was treated in the same tolerant fashion, although it is not so easy, of course, for a conscientious examiner to discover two extra marks as it is to discover one; and if the candidate is 3 marks below the pass line it is still harder to give him the necessary lift to carry him over. Thus in the final list we find more condidates with 26 marks than with 27, and still more with 25 than with 26. If the above diagnosis is correct, and all marks-sheets tell the same tale, who shall again say that examiners do not temper justice with mercy?

This example has illustrated fairly clearly the difficulty of fixing the mode with any great precision by mere inspection when the individuals are arranged in groups, the value of the variable under discussion lying between prescribed limits for each group. While

it is possible to get a rough approximation to its value in this way, we conclude that for a really satisfactory determination we require some method which makes use of the whole distribution, as in the determination of the mean, and not merely of the portion in the supposed neighbourhood of the mode. This must be left to a later chapter; we shall only point out before passing on that there may sometimes be more than one mode in a given frequency distribution just as there may be more than one fashionable type of collar which it is expedient for the draper to stock in large quantities. The second grouping in the examination example suggests such a possibility, for it will be noticed that the frequencies of candidates do not rise steadily to a single maximum at 108 for class (28 to 32), and then fall steadily: there is a previous rise and fall in the neighbourhood of class (18 to 22).

Weighted Mean. Let us suppose a farmer employs for the harvest 5 men, 3 women, and 4 boys. In estimating the amount of work they can do in a given time it is clear that in general a woman or boy cannot be reckoned as equal to a man. He must therefore decide what ' weight ' must be given to each in proportion to a man. If a woman's work be taken, for example, to be three-quarters as effective and a boy's work to be half as effective as that of a man, we have as the appropriate proportional weights

$$1 : \tfrac{3}{4} : \tfrac{1}{2}, \text{ or } 4 : 3 : 2.$$

Hence 5 men, 3 women, and 4 boys would on the average be equivalent in output to

$$\left(5+3\times\tfrac{3}{4}+4\times\tfrac{1}{2}\right) \text{ men}$$

$$=\frac{4\times5+3\times3+2\times4}{4} \text{ men}$$

$$=9\tfrac{1}{4} \text{ men.}$$

An average of this type is called a weighted mean, 1, $\frac{3}{4}$, and $\frac{1}{2}$ being the weights, because they tell us what weight to give to each separate worker in calculating the average.

Let us consider the effect such weighting has in general upon a mean, and for this purpose we shall test it on a set of index numbers measuring rents in certain groups of towns in 1912, as given in a *Report on the Cost of Living of the Working Classes* issued by the Board of Trade (Cd. 6955).

TABLE (9). MEAN INDEX NUMBERS OF RENTS FOR CERTAIN GEOGRAPHICAL GROUPS OF TOWNS IN 1912 (WITH REFERENCE TO MIDDLE ZONE OF LONDON AS STANDARD = 100).

(1) Geographical Group.	(2) Rents.	(3) No. of Towns included in the Group.	(4) Each Group counting as 1.	(5) Arbitrary Weights.	(6) Approximate sub-multiples of Nos. in previous column.
Northern Counties and Cleveland	66·0	9	1	27	3
Yorkshire (except Cleveland)	58·5	10	1	54	6
Lancashire and Cheshire .	56·9	17	1	45	5
Midlands	52·3	14	1	125	14
Eastern and East Midland Cos.	53·4	7	1	63	7
Southern Counties . .	63·7	10	1	14	2
Wales and Monmouth . .	64·8	4	1	22	2
Scotland	62·0	10	1	178	20
Ireland	51·7	6	1	55	6
Average . . .	..	58·4	58·8	57·6	57·6

The first mean in the above table, 58·4, is obtained by multiplying (or weighting) the mean rent of each geographical group by the number of towns in the group, given in col. (3), adding the numbers so obtained, and dividing the total by the total number of towns, thus :—

$$\frac{9(66{\cdot}0)+10(58{\cdot}5)+\ .\ .\ .\ +6(51{\cdot}7)}{9\ +\ 10\ +\ .\ .\ .\ +\ 6}.$$

This is simply the arithmetic mean treating each town as unit.

The second mean, 58·8, is obtained by adding the mean rents of all the groups and dividing by the total number of groups, thus :—

$$\frac{66{\cdot}0+58{\cdot}5+\ .\ .\ .\ +51{\cdot}7}{1\ +\ 1\ +\ .\ .\ .\ +\ 1}.$$

This is the arithmetic mean treating each geographical group as unit.

The third mean, 57·6, is obtained by multiplying, or weighting, the mean rent of each group by a perfectly arbitrary number given in col. (5) ; the numbers selected were taken quite at random from

another column of figures in another Blue-book, and had no connection whatever with the subject of rents ; this gives :—

$$\frac{27(66 \cdot 0)+54(58 \cdot 5)+ \ . \ . \ . \ +55(51 \cdot 7)}{27 \ + \ 54 \ + \ . \ . \ . \ + \ 55}.$$

The last mean, 57·6, is obtained by choosing as weights any numbers (and for simplicity we choose the smallest) as in col. (6) which are very roughly proportional to the arbitrary weights used in the last instance ; we thus get :—

$$\frac{3(66 \cdot 0)+6(58 \cdot 5)+ \ . \ . \ . \ +6(51 \cdot 7)}{3 \ + \ 6 \ + \ . \ . \ . \ + \ 6}.$$

Now the first of these means is clearly the most satisfactory, since it is the result of very properly weighting the mean rent of each group of towns according to the number of towns the group contains. But the second result shows that if we are ignorant of the number of the towns in each group we shall not be very far out in our calculation if we treat them all as of equal importance, and find the simple arithmetic mean of the mean rents in the nine groups. We can even go further, for we find, from the third and fourth results, that by weighting the mean rents in the various groups on quite a random basis, the mean we get still does not differ very greatly from the best value first found.

The important principle of which the above example is an illustration is perfectly general, and may be stated as follows: If the total number of measurements or observations be not very small, and if the resulting values of the organ or character measured (rent in our case) be not very unequal, any reasonable selection of multipliers or weights (as, for instance, the first two adopted above) will give means which differ from one another by but little ; and even an apparently unreasonable selection of multipliers (as, for instance, the third adopted above), assuming they are not so wildly chosen as to give any particular group a very unfair weight in comparison with the others, will not throw the mean out badly. Further, in place of a set of large multipliers we may substitute small numbers which are roughly proportional to them (as we have done in the fourth case above), and the mean will again be very little affected. [See Appendix, Note 2.]

CHAPTER V

AVERAGES (*continued*)

Applications of Weighted Mean. In determining the weighted mean of a set of observations it is usual, of course, to weight each observation according to its importance, though what number should be chosen as a measure of its importance may sometimes be a matter of doubt. It is not a very difficult matter to decide when we wish, for example, to compare birth, marriage, or death rates in two districts, if we know how the constitution of the population in the one district differs from that in the other, for the weighting in each of these cases must be in proportion to the population concerned, and it is too important to ignore.

Death rate, crude and corrected. Imagine a city in which the total number of deaths in a certain year is N out of a population numbering P.

The ordinary or crude death rate for that city will then be $\frac{N}{P}\times 1000$, by definition.

Now this number N may be analysed according to the ages of the people who have died; let us suppose it is made up of

n_1	people	between limits	0	and less than	5	years of age,
n_2	,,	,,	,, 5	,,	15	,,
n_3	,,	,,	,, 15	,,	25	,,

and so on, where

$$n_1+n_2+n_3+\ .\ .\ .\ =N.$$

Again the number P may be analysed according to the ages of the people who compose the total population, giving, say,

p_1	of the population	between limits	0	and less than	5	years of age,
p_2	,,	,,	,, 5	,,	15	,,
p_3	,,	,,	,, 15	,,	25	,,

and so on, where

$$p_1+p_2+p_3+\ .\ .\ .\ =P.$$

Thus we may write for the crude death rate

$$D=\frac{N}{P}\times 1000$$

$$=\frac{n_1+n_2+n_3+\ .\ .\ .}{P}\times 1000$$

$$=\frac{n_1}{P}1000+\frac{n_2}{P}1000+\frac{n_3}{P}1000+\ .\ .\ .$$

$$=\frac{p_1}{P}\left(\frac{n_1}{p_1}1000\right)+\frac{p_2}{P}\left(\frac{n_2}{p_2}1000\right)+\frac{p_3}{P}\left(\frac{n_3}{p_3}1000\right)+\ .\ .\ .$$

$$=(p_1d_1+p_2d_2+p_3d_3+\ .\ .\ .)/P,$$

where d_1 is the death rate between limits 0 and less than 5 years of age,

d_2	,,	,,	,, 5	,,	15	,,
d_3	,,	,,	,, 15	,,	25	,,

and so on.

Now if we compare this expression with the corresponding one for another city, say,

$$D'=(p'_1d'_1+p'_2d'_2+p'_3d'_3+\ .\ .\ .)/P',$$

it is quite conceivable that the death rates in the various age groups might be equal—

$$d_1=d'_1,\ d_2=d'_2,\ d_3=d'_3\ .\ .\ .$$

and yet D might exceed D′ because in the first city there are a greater proportion of infants or old people, on which classes the hand of death falls heaviest, that is, because the p's or weights which multiply the biggest d's are greater in the first case than in the second. But so long as the d's in the two cities are equal, age group by age group, it would be reasonable to regard the cities as equally healthy, or unhealthy as the case might be, and therefore to insure a fair comparison it is usual in the Reports of the Registrar-General to give a corrected death rate in place of the crude death rate defined above.

This is done by weighting the death rate for each age group, not in proportion to the actual number of persons in that group in the city itself, but in proportion to the corresponding number in

the country at large. Thus, if we denote the proportion of the population, Q,

between limits	0	and less than	5	in the country	at large	by	q_1/Q,
„	„ 5	„	15	„	„	„	q_2/Q,
„	„ 15	„	25	„	„	„	q_3/Q,

and so on, we get as the corrected death rate

$$(q_1d_1+q_2d_2+q_3d_3+\ .\ .\ .)/Q,$$

a form which has the effect of making the results agree in two cities which have equal d's throughout.

A similar method of correction is clearly applicable in considering the incidence of the death rate when we are concerned not with a difference of district but with a difference of sex, occupation, religious profession, wage-earning capacity, or any other well-defined character. Further, it may be used also in comparing birth rates, marriage rates, heights, weights, chest measurements, or any similar attributes, when it is necessary to refer the observations or measurements to a standard population in order to avoid complications due to age variation.

There is another method of correction, equally general in application, which is useful when the death rates in the various age groups are not known. In this case D, the crude death rate for the whole population of the district is known, also p_1/P, p_2/P, p_3/P, . . . the proportions of the population between the various age limits, but d_1, d_2, d_3 . . . are supposed unknown.

Now if the population in the country as a whole were the same in corresponding age groups as it is in the district under consideration, we should get as the death rate for the whole country

$$(p_1\delta_1+p_2\delta_2+p_3\delta_3+\ .\ .\ .)/P,$$

where δ_1, δ_2, δ_3 . . . are the death rates in the various age groups in the country at large, and these would in practice as a rule be known.

The actual death rate for the whole country is, however,

$$(q_1\delta_1+q_2\delta_2+q_3\delta_3+\ .\ .\ .\ )/Q,$$

where q_1/Q, q_2/Q, q_3/Q . . . denote, as before, the real proportions of the population in the various age groups in the country at large.

We take as the corrected death rate required for the district a number bearing to the crude death rate the same ratio as

$$(q_1\delta_1+q_2\delta_2+\ .\ .\ .)/Q \text{ bears to } (p_1\delta_1+p_2\delta_2+\ .\ .\ .)/P.$$

Hence we have

$$\frac{\text{corrected death rate}}{\text{D}}=\frac{q_1\delta_1+q_2\delta_2+\ .\ .\ .}{p_1\delta_1+p_2\delta_2+\ .\ .\ .}\times\frac{\text{P}}{\text{Q}}.$$

Index Numbers to compare Household Budgets. Another highly important illustration of a weighted mean occurs in the search for a satisfactory measure of the change in the cost of living from year to year. We have already introduced the subject of variation in wholesale prices, and we have seen that Sauerbeck, in forming his index numbers, treats as one each of the forty-five commodities he uses to measure this variation: the observations, that is to say, are not weighted.

But, confining our attention to food alone, supposing we have five items, such as bacon, bread, tea, sugar, milk, for which the index numbers of prices at two different dates are :—

	Bacon.	Bread.	Tea.	Sugar.	Milk.
First date	100	100	100	100	100
Second date	117	95	94	102	109

Is it really right to treat each of these items as of equal importance with the rest, or ought we to regard bread and tea, say, as of more weight than bacon, and count bread perhaps five times and tea three times while counting bacon only once ? It is clear that, in order to select a reasonable set of multipliers in this case, we should need to know the standard of living of the class of people under consideration, and how much in the aggregate they spend upon bacon and how much upon bread, etc.

A partial answer to these questions can be obtained by making a collection of household budgets as was done, for example, by two Government Committees which recently reported (1918-19) on the *Cost of Living among the Urban and the Agricultural Working Classes* respectively. If the number of commodities employed is large, even an arbitrary set of multipliers, as we have indicated, will not displace the mean any great distance from the value when reasonable weights are chosen, but unfortunately in collecting such household budgets we are confined to the comparatively limited variety of food-stuffs which are in general use.

Different principles may be followed in making the comparison

between one year and another which may be illustrated by a few figures from the Urban Classes Report (1918) :—

TABLE (10). HOUSEHOLD BUDGETS SHOWING PRICES OF EACH COMMODITY AND QUANTITIES PURCHASED AT TWO DIFFERENT DATES BY TYPICAL FAMILY.

Commodity.	First year (1914).		Second year (1918).	
	x_1 Price (pence per lb).	n_1 No. of lb. bought.	x_2 Price (pence per lb.)	n_2 No. of lb. bought.
Sugar .	2·2	5·9	7·07	2·83
Tea . .	21·3	0·68	33·3	0·57
Potatoes .	0·7	15·6	1·25	20·0
..	..	..	..	..
..	..	..	..	..

Let x_1 be the price, in pence per unit, of any one commodity at the first date, and let n_1 be the number of units of this commodity bought per week by a typical family (n may be estimated in different ways, *e.g.* (1) by dividing the total number of units bought by all families by the total number of those families, or (2) by ranging the different amounts bought by different families in order of magnitude and picking out the median amount, or (3) by choosing the mode, *i.e.* the amount most commonly purchased). Also let x_2 be the price, in pence per unit, of the same commodity at the second date, and let n_2 be the number of units of the commodity then bought per week by the typical family estimated in the same way as before.

The actual expenditure, measured in pence, at the two dates will then be

$$\Sigma(x_1n_1) \text{ and } \Sigma(x_2n_2)$$

respectively, where $\Sigma(x_1n_1)$ simply denotes the sum of expressions like (x_1n_1) for all the commodities recorded and $\Sigma(x_2n_2)$ denotes the sum of expressions like (x_2n_2) for all the commodities recorded, Σ, the old English S, being a well-known conventional abbreviation for ' Sum of expressions like.' Thus, with the numbers in Table (10), we should have

$$\Sigma(x_1n_1)=(2{\cdot}2)(5{\cdot}9)+(21{\cdot}3)(0{\cdot}68)+(0{\cdot}7)(15{\cdot}6)+\ \ldots$$
$$\Sigma(x_2n_2)=(7{\cdot}07)(2{\cdot}83)+(33{\cdot}3)(0{\cdot}57)+(1{\cdot}25)(20{\cdot}0)+\ \ldots$$

Taking 100 as the index number to represent expenditure at the first date, the index number measuring expenditure at the second date may be formed in any of the following different ways,* which as a rule, of course, lead to different results :—

(1) $100\Sigma(x_2n_2)/\Sigma(x_1n_1)$;
(2) $100\Sigma(x_2n_1)/\Sigma(x_1n_1)$ or $100\Sigma(x_2n_2)/\Sigma(x_1n_2)$;
(3) $100\Sigma(x_1n_2)/\Sigma(x_1n_1)$ or $100\Sigma(x_2n_2)/\Sigma(x_2n_1)$.

The first of these expressions compares the *actual expenditure* at the second date to that at the first date.

The next two expressions take into account directly only the change in prices ; they compare, not actual expenditures but, the expenditures at the two dates as they would be if the amounts purchased at the two dates were the same : the first supposing these amounts to equal those actually bought at the first date, and the second supposing them to equal those actually bought at the second date.

The last two expressions, on the other hand, take into account directly only the change in amounts purchased ; they compare the expenditures at the two dates as they would be if the prices ruling at the two dates were the same : the first supposing these prices to equal those actually charged at the first date, and the second supposing them to equal those actually charged at the second date.

The particular method of weighting adopted must naturally depend upon the circumstances of the period under discussion and the nature of the inquiry one is making ; it is a nice question to decide how far emphasis should be laid upon the old standard of life (measured by food, lighting, rent, recreation, etc.) with the expense required to maintain it, and upon the new standard of life and the cost necessary to reach it.

It may be useful here to summarize a few of the questions of interest which present themselves in connection with the formation of index numbers of prices designed to measure changes in the value of money in general without reference to any particular class of the community :—

1. What years should be selected in fixing our standard prices ?

2. What commodities should be chosen as a basis for our average ?

[* See also *The Measurement of Changes in the Cost of Living*, by A. L. Bowley, Sc.D., in the *Journal of the Royal Statistical Society*, May 1919, for a more complete discussion of the subject.]

3. What weight should be given to each commodity in relation to the rest ?

4. How should the prices of the several commodities be determined, bearing in mind that 'price' itself frequently varies from place to place ?

5. Finally, how should these prices be combined to give the average required ? Should we use the simple arithmetic mean, the geometric mean [see Appendix, Note 3], the median, or some other measure ?

While we are not prepared to attempt to answer these questions fully, seeing that authorities are not altogether agreed as to what the answers should be, one or two points may be worth noting. Generally speaking we may say that :—

1. The years selected in fixing our standard prices should be years in which economic conditions were normal rather than abnormal.

2. The commodities chosen should be articles of general consumption, and as wide a field as possible should be covered in their choice.

3. Many consider that little is gained by weighting, but, if weights are introduced, the greater the importance of any commodity in relation to the rest, judged for example by the relative quantity consumed, the greater should be the weight assigned to it.

4. The practical difficulty of assessing retail prices when they are uncontrolled compels us in general to fall back upon wholesale quotations, on which some light may be thrown by keeping under observation the important markets for the sale of each commodity.

5. The average commonly used is the simple arithmetic or the weighted mean, though arguments can be adduced in favour of other averages such as the median.

Leaving index numbers now on one side and returning to the general subject of averages, we may remark that the question which average is correct in any given case, the mean (weighted or otherwise), the median, or the mode, does not arise : no one average is more correct than another, because they are all entirely conventional and represent different ideas ; they correspond in fact to so many different ways of summing up a set of observations or measurements in a single numerical statement, and the real question

to determine is which statement, which kind of average, brings the set of observations before us to the best focus.

For this purpose one average will clearly be best in one case and another in another, but it may be stated without hesitation that the arithmetic mean is certainly the most useful of the three and it is the most frequently used. Other averages have been suggested, such as the *geometric* and the *harmonic means* [see Appendix, Note 3] familiar to students of Algebra, as suitable in special classes of problems.*

In a reasonably symmetrical distribution of observations, one in which the variables of medium size are the most frequent and the frequency diminishes about equally on either side towards the largest and the least of the variables, the values of the mean, the median, and the mode will be found to lie all very close together; and a useful practical rule to remember is that *the median comes in general between the mean and the mode, the difference between the mean and the mode being about three times the difference between the mean and the median.* This rule, for lack of a better, might be used to determine the mode in suitable cases, or it might be used to test the value found in some other way.

The general term 'average' is frequently used when the particular denomination 'arithmetic mean' is implied, but the context will usually prevent misunderstanding.

In order to get a clear impression of the outstanding features presented by the three chief averages discussed, let us go over them once more in the case of marks awarded to a number of students in a class. All three may be regarded as in a sense measures of the standard reached by the class as a whole in the examination, but the measures are made in different ways :—

1. *The Arithmetic Mean* is found by merely dividing the aggregate marks of the class by the number of the students, and it gives the marks earned by each student if we conceive them all to be of equal merit.

2. *The Median* is found by ranging the students in order of merit from top to bottom, and picking out the marks awarded to the one who comes half-way down the list.

3. *The Mode* is the most fashionable number of marks, *i.e.* the marks obtained by the greatest number of candidates.

The advantages and disadvantages of the three types may be set out broadly as follows, although the boundary lines must not be too strictly drawn :—

* See Note on p. 41.

Mean.	Median.	Mode.
Easy to calculate when the values of the variable can be summed and their number is known.	Easy to pick out when the individuals can be ranged in order according to the value or degree of the variable observed.	Not easy to determine with precision, when the observations fall into groups of different ranges, without fitting a frequency curve to the distribution as a whole.
Well designed for algebraical manipulation, as, for example, when we wish to combine different sets of observations [see Appendix, Note 4, for two illustrations].	Unsuited for algebraical work.	Unsuited for algebraical work.
Affected sometimes too much by abnormal individuals among the observations.	Determined merely by its position in the distribution, and its actual value is thus quite unaffected by abnormal individuals.	Unaffected by abnormal individuals, and owes its importance to the fact that it is located in the region where the frequency is most dense.

The reader should test his grasp of the principles so far introduced by applying them himself to a concrete case. For example, he might use the data in Table (11), with regard to wages earned by certain women, taken from Tawney's *Minimum Wages in the Tailoring Trade*, and based upon the 1906 Wages Census. Let him begin by roughly estimating the mean, the median, and the mode from an inspection of the distribution. He might then proceed to calculate the mean wage :—

(1) taking the actual frequencies given in the table ;
(2) taking simple sub-multiples of these frequencies, roughly one-hundredth part of each : 2, 4, 6, 7, 9, 11, etc. ;
(3) assuming unit frequency in place of that given in the table for each wage group.

Finally, he might determine the median and the mode in the manner explained in the text, deducing the latter from the relation (mean−mode)=3(mean−median).

The results obtained should be

(1) 13·08s.; (2) 13·10s.; (3) 15·59s.
Median = 12·53s.; Mode = 11·43s.

TABLE (11), DISTRIBUTION OF WAGES OF CERTAIN WOMEN TAILORS.

(1) Wages between limits	(2) No. of Women earning wages as shown in Column (1).	(3) Wages between limits	(4) No. of Women earning wages as shown in Column (3).
5s. and less than 6s.	180	16s. and less than 17s.	642
6s. ,, ,, 7s.	384	17s. ,, ,, 18s.	453
7s. ,, ,, 8s.	553	18s. ,, ,, 19s.	401
8s. ,, ,, 9s.	690	19s. ,, ,, 20s.	272
9s. ,, ,, 10s.	900	20s. ,, ,, 21s.	251
10s. ,, ,, 11s.	1145	21s. ,, ,, 22s.	138
11s. ,, ,, 12s.	1201	22s. ,, ,, 23s.	124
12s. ,, ,, 13s.	1138	23s. ,, ,, 24s.	64
13s. ,, ,, 14s.	930	24s. ,, ,, 25s.	54
14s. ,, ,, 15s.	885	25s. ,, ,, 30s.	122
15s. ,, ,, 16s.	790	.. ..	..

* [One important example of the use of the geometric mean is in the construction of the Board of Trade Index Number of Wholesale Prices—see article by A. W. Flux, C.B., M.A., in the *Journal of the Royal Statistical Society*, March 1921.]

CHAPTER VI

DISPERSION OR VARIABILITY

LET us suppose that two men set out separately on walking tours and that they walk as follows :—

	First Man walks	Second Man walks
First day . . .	20 miles.	15 miles.
Second „ . . .	20 „	20 „
Third „ . . .	25 „	25 „
Fourth „ . . .	25 „	25 „
Fifth „ . . .	30 „	30 „
Sixth „ . . .	30 „	35 „
6 days . . .	150 miles.	150 miles.

The total distance covered in six days, namely 150 miles, and therefore also the mean rate of walking, 25 miles a day, are thus exactly the same in both cases, but the *dispersion* of the values of the variable (the variable being in this instance the number of miles walked per day) round about their mean value, the *variability*, is different in the two cases. The greatest deviation from the average in the first case is five and in the second case it is ten miles.

Thus, besides knowing the average of a set of values of a variable it is important to measure the dispersion of the distribution. Are the observations crowded in a dense mass around the average, or do they tail off above and below it, and to what extent ? In other words, what is the variability from the average of the distribution ?

Mean Deviation. Now we are not concerned here with the signs of the separate deviations, with the question, that is, whether any particular value of the variable lies above or below the average :

it is only of their amount we wish to take cognizance, and perhaps the most obvious way to measure the total variability and at the same time to ignore the signs of the separate deviations from the average is to add up these deviations, treating them all as signless, and to divide the result by their total number. This gives what is known as the mean deviation of the system of observations—it is the ordinary arithmetic mean of the separate deviations, treated as if they are all in the same direction, and, in measuring them, we may use either the mean or the median as the average, but it would seem preferable to take the latter because the mean deviation is least when the median is chosen as the origin, or zero point, from which the differences are measured. The proof of this fact will be found in Note 6 in the Appendix, but we may readily test it in a given case.

Let us adapt the 'walking' illustration used above, slightly extending the figures and making them unsymmetrical, *i.e.* of unequal variability on either side of the average, so as to prevent the median coinciding with the mean. We then have an amended table setting out the number of miles walked by a certain man on successive days during, say, a fortnight's tour, as follows :—

TABLE (12). NUMBER OF MILES WALKED ON SUCCESSIVE DAYS.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
f No. of days.	Miles walked.	x Deviation from 25.	x_1 Deviation from 24·64.	x_2 Deviation from 24.	x_3 Deviation from 26.	fx [No. in Col. (1)]×[No. in Col. (3)].	fx_1 [No. in Col. (1)]×[No. in Col. (4)].
1	10	15	14·64	14	16	15	14·64
2	15	10	9·64	9	11	20	19·28
3	20	5	4·64	4	6	15	13·92
3	25	..	0·36	1	1	..	1·08
2	30	5	5·36	6	4	10	10·72
2	35	10	10·36	11	9	20	20·72
1	40	15	15·36	16	14	15	15·36
14	..	..	..	..	..	95	95·72

The first two columns show that 10 miles was the distance walked on the first day, 15 miles on each of the next two days, 20 miles on each of the next three days, and so on until the last day, when 40 miles was the distance walked.

The median in this case, being the number of miles walked on the middle day when the days are ranged in order of mileage from the least to the greatest, is 25, for this is the distance covered on both the seventh and the eighth days which come half-way along the series.

Col. (3) shows the deviations from the median, 25, of the distances covered each day as recorded in col. (2), and col. (7) enables us to sum these deviations when each is multiplied by the number of days to which it corresponds, since these numbers, given in col. (1), show how many times each deviation is repeated. Hence the mean deviation, regardless of sign, measured from the median

$$
\begin{aligned}
&=[(1\times 15)+(2\times 10)+(3\times 5)+(2\times 5)+(2\times 10)+(1\times 15)]/14\\
&=(15+20+15+10+20+15)/14\\
&=95/14\\
&=6{\cdot}79 \text{ miles.}
\end{aligned}
$$

We may compare this with the corresponding deviations measured from (1) the arithmetic mean, (2) the number 24, and (3) the number 26 as origin respectively.

1. The arithmetic mean of the distribution is obtained at once by multiplying the corresponding numbers in cols. (1) and (2), adding the results, and dividing the total by 14, thus

$$
\begin{aligned}
\text{Arithmetic mean} &= \frac{1(10)+2(15)+3(20)+3(25)+2(30)+2(35)+1(40)}{1+2+3+3+2+2+1}\\
&= \frac{10+30+60+75+60+70+40}{14}\\
&= 345/14\\
&= 24{\cdot}64 \text{ miles,}
\end{aligned}
$$

and the deviations from 24·64 are shown in col. (4); the mean deviation from 24·64, obtained by combining cols. (1) and (4) and adding as shown in col. (8)

$$
\begin{aligned}
&=[1(14{\cdot}64)+2(9{\cdot}64)+\ \ldots\ ]/14\\
&=95{\cdot}72/14\\
&=6{\cdot}84 \text{ miles.}
\end{aligned}
$$

2. Similarly, the mean deviation from 24, making use of col. (5),

$$
\begin{aligned}
&=[1(14)+2(9)+\ \ldots\ ]/14\\
&=6{\cdot}93 \text{ miles.}
\end{aligned}
$$

3. And the mean deviation from 26, making use of col. (6),

$$=[1(16)+2(11)+ \ . \ . \ . \ ]/14$$
$$=7{\cdot}07 \text{ miles.}$$

The original determination gives a value which is less than any of these three results, as was anticipated.

The mean deviation from the median is, however, difficult to calculate with exactness when the observations are recorded in groups between different limits: for this and other reasons we shall not spend much time upon it, and we shall as a rule choose the mean as origin of reference rather than the median. It may be as well to explain the source of the difficulty by a small hypothetical illustration.

Let us suppose that in making measurements of some organ or character in 13 individuals we get a result lying between 4 and 6 units on six occasions, between 6 and 8 units on four occasions, and between 8 and 10 units on three occasions. Here, *assuming* that all the individuals in any group have the mid-value measurement for that group, *i.e.* treating the distribution as one of 6 individuals with a variable measuring 5 units, 4 individuals with a variable measuring 7 units, and 3 individuals with a variable measuring 9 units, we get $\frac{18}{13}$ as the mean deviation with 7 as origin and $\frac{18\frac{1}{2}}{13}$ for the mean deviation with 6·5 as origin, as the following table shows :—

Measurement.	f Frequency.	x Deviation from 7.	y Deviation from 6·5.	fx	fy
4 and less than 6	6	2	1·5	12	9
6 „ „ 8	4	0	0·5	..	2
8 „ „ 10	3	2	2·5	6	7·5
	13	..	..	18	18·5

Now the result obtained is in agreement with the minimum mean deviation theory, granted that 7 is the median measurement, as it might certainly be. But it is not so of necessity, and in that case the assumption italicized might lead, in the above calculation, to appreciable inaccuracy unless the number of observations is large and the class-interval is small. For example, the actual

distribution might, without contradicting the previous data, conceivably run :—

Measurement.	f' Frequency.	x' Deviation from 7.	y' Deviation from 6·5.	$f'x'$	$f'y'$
5	6	2	1·5	12	9
6·5	2	0·5	..	1	..
7·5	2	0·5	1	1	2
9	3	2	2·5	6	7·5
..	13	..	..	20	18·5

But in this case the median, the measurement for the seventh individual from either end of the series, is 6·5, and according to the first calculation the mean deviation referred to 6·5 as origin appears to be greater than that referred to 7 as origin. If, however, we recalculate, using the more detailed table, we find that the mean deviation referred to 6·5 as origin ($\frac{18\frac{1}{2}}{13}$) is really less than the mean deviation with reference to 7 as origin, as it should be, for the latter now turns out to be $\frac{20}{13}$.

Standard Deviation. An alternative method of avoiding the signs of the deviations from the average in order to estimate the amount of variability of the distribution is to square each separate deviation, sum the squares, divide by their number, and take the square root of the result. This gives the *root-mean-square deviation*, and it is least when the arithmetic mean of the variables is chosen as origin from which to measure the deviations, when it is known as the standard deviation. For proof of this minimum principle see Appendix, Note 5, but it is worth while testing it also with the data given in Table (12).

The numbers in cols. (3) to (6) in Table (13) are obtained simply by squaring the corresponding numbers in the same cols. (3) to (6) in Table (12). Col. (7) is formed in order to enable us to calculate the mean-square deviation referred to 25 as origin; the numbers in col. (3) show the squares of the deviations for each individual observation, and the numbers in col. (1), by which they are multiplied, show how frequently the same values are repeated. Hence we get the mean-square deviation with reference to 25

$$=[1(225)+2(100)+3(25)+2(25)+2(100)+1(225)]/14$$
$$=975/14$$
$$=69{\cdot}64.$$

Thus the root-mean-square deviation referred to 25

$$=\sqrt{(69\cdot64)}$$
$$=8\cdot345.$$

Similarly, by means of col. (8), formed on exactly the same principle, we find that the root-mean-square deviation referred to 24·64 as origin

$$=\sqrt{[(214\cdot33+185\cdot86+\ .\ .\ .\ )/14]}$$
$$=\sqrt{(973\cdot22/14)}$$
$$=8\cdot338.$$

But 24·64 is the mean of the distribution, hence 8·338 is the standard deviation.

With the help of cols. (5) and (6) the student may himself calculate the root-mean-square deviation with regard to 24 and 26 respectively as origin; the results should be 8·36 and 8·45. Of the four values thus obtained for the root-mean-square deviation, the least is that referred to the mean as origin, the standard deviation, now proposed as a measure of variability or dispersion suitable for most general purposes.

This measure possesses several decided advantages over the mean deviation; among others it lends itself more easily to certain algebraical processes (see, for example, p. 158), a fact of importance when we wish, for instance, to discuss two sets of observations in combination, and it is in general less affected by 'fluctuations of sampling'—errors which arise owing to the fact that we cannot as a rule survey the whole field of operations, but have to be content with a sample.

TABLE (13). NUMBER OF MILES WALKED ON SUCCESSIVE DAYS.

(1) f No. of days.	(2) Miles walked.	(3) x^2 Square of Deviation from 25.	(4) x_1^2 Square of Deviation from 24·64	(5) x_2^2 Square of Deviation from 24.	(6) x_3^2 Square of Deviation from 26.	(7) fx^2 [No. in Col (1)] × [No. in Col. (3)]	(8) fx_1^2 [No. in Col. (1)] × [No. in Col. (4)]
1	10	225	214·33	196	256	225	214·33
2	15	100	92·93	81	121	200	185·86
3	20	25	21·53	16	36	75	64·59
3	25	..	0·13	1	1	..	0·39
2	30	25	28·73	36	16	50	57·46
2	35	100	107·33	121	81	200	214·66
1	40	225	235·93	256	196	225	235·93
14	..	..	..	..	..	975	973·22

Quartile Deviation or Semi-interquartile Range. There is a third measure of dispersion, based upon the determination of the *quartiles*, and to introduce them we may refer again to Table (7) in order to show how the idea of the median may be extended.

We define the individual occupying a position one-quarter the way along any series of observations, arranged in ascending order of magnitude of some organ or character common to all the individuals of the series, as the *lower quartile*; and we define the individual occupying a position three-quarters the way along the series as the *upper quartile.*

When the distribution of observations is divided up into groups lying between different limits of the variable under consideration the quartiles may, like the median, be calculated by interpolation. Thus, in the examination example, the total number of candidates is 514 and $\frac{1}{4}(514)=128{\cdot}5$.

But the 91st candidate from the bottom gets approximately 20 marks, and the 149th candidate from the bottom gets approximately 25 marks. Hence the imaginary candidate, No. 128·5, should get a number of marks lying somewhere between 20 and 25. But if, in this neighbourhood, a difference of

(149-91) candidates corresponds to a difference of 5 marks,

(128·5-91) ,, should correspond ,, $5\times\frac{37{\cdot}5}{58}$ marks.

Thus, the marks assigned to the lower quartile candidate are approximately

$$=20+\frac{5\times 37{\cdot}5}{58}$$

$$=20+3{\cdot}23.$$

Hence *the lower quartile*$=23{\cdot}23$.

Again $\frac{3}{4}(514)=385{\cdot}5$.

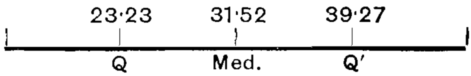
But the 318th candidate from the bottom gets approximately 35 marks, and the 397th candidate from the bottom gets approximately 40 marks. Therefore, the imaginary candidate, No. 385·5, should get approximately a number of marks

$$=35+5\times\frac{67{\cdot}5}{79}$$

$$=39{\cdot}27.$$

Hence *the upper quartile*$=39{\cdot}27$.

It is clear that the quartiles together with the median divide the whole series of observations into approximately four equal groups, so that the quartile marks give a rough idea of the distribution on either side of the average. For this reason half the difference between the quartiles provides a convenient measure of the dispersion, and it is called the quartile deviation or semi-interquartile range ; thus, if Q be the lower and Q′ the upper quartile, we have



$$\textit{the quartile deviation} = \tfrac{1}{2}(Q' - Q).$$

In the above example, this measure

$$= \tfrac{1}{2}(39{\cdot}27 - 23{\cdot}23)$$
$$= \tfrac{1}{2}(16{\cdot}04)$$
$$= 8{\cdot}02.$$

If a more minute analysis of the distribution of variables is desired, we may range them in order of magnitude as before, and divide up the series into ten equal parts, recording every tenth along the line ; these tenths are called *deciles.*

Thus, the deciles in the examination example correspond to the marks assigned to imaginary candidates numbered as follows :—

51·4, 102·8, 154·2, 205·6, 257·0, 308·4, 359·8, 411·2, 462·6,

and they can be calculated by the interpolation method used in finding the median and quartiles.

This way of representing the chief features of a distribution, by quartiles, etc., was much used by Galton in his researches and writings.

The student may be perplexed as to which should be used of so many different measures of dispersion or variability, but there need be no real confusion. If a rough estimate only is wanted the quartile deviation is a convenient measure, assuming that the variables observed or measured can be ranged in order of magnitude so as to admit of the quartiles being readily picked out. Also the measure thus obtained is not unsatisfactory when the distribution of values of the variable is fairly symmetrical and uniform in its gradation from greatest frequency to least. If, however, it is conspicuously skew (unsymmetrical) and there are erratic differences in frequency between successive values of the variable, it is better to choose a measure which gives the magnitude and the position of each recorded observation its due weight in the deviation sum.

Then again the choice as between the standard deviation and the mean deviation may be sometimes determined by the particular kind of average which suits the problem best. But as the arithmetic mean is the most important and the most commonly used average, so the standard deviation is certainly the most important measure of dispersion.

It will be shown later that the following relations are *approximately* true when the distribution of variables is not very far from being symmetrical :—

(1) *Quartile deviation* $=\frac{2}{3}$(*Standard deviation*).
(2) *Mean deviation* $=\frac{4}{5}$(*Standard deviation*).

In (2) the mean deviation should be measured from the mean.

Also (3) a range of two or three times the standard deviation on both sides of the mean will be found to include the majority of the observations in the distribution.

Coefficient of Variation. Before we pass on to illustrate the subject of averages and variability by means of a few examples it is necessary to introduce one more constant known as the coefficient of variation. It is a measure of variability but it differs from the chief measures already discussed in that they are absolute measures, whereas the coefficient of variation, written C. of V. for short, is a ratio or relative measure. The need for it arises when we reflect that in order to gauge fairly the amount of variability we ought to have in mind also the size of the mean from which the variation is measured; just as a difference of 1 foot between the heights of two men is a conspicuous difference when the normal height is between 5 and 6 feet, whereas the same difference of 1 foot between two measured miles would be trifling because the standard mile contains over 5000 feet.

The coefficient of variation has been defined by Karl Pearson (*Phil. Trans.*, vol. 187A p. 277), who first suggested its use, as 'the percentage variation in the mean, the standard deviation (S.D.) being treated as the total variation in the mean,' so that

C. of V. = 100 S.D./Mean.

He pointed out that it would be idle, in dealing with the variation of men and women (or indeed very often of the two sexes of any animal), to compare the absolute variation of the larger male organ directly with that of the smaller female organ, because several of these organs, as well as the height, the weight, brain capacity, etc.,

are greater in man than in woman in the approximate proportion of 13 : 12.

As an example of the use of the C. of V., figures may be quoted from a paper by R. Pearl and F. J. Dunbar (*Biometrika*, vol. ii. pp. 321 *et seq.*), *On Variation and Correlation in Arcella*. Measurements in mikrons were made of the outer and inner diameters of 504 specimens of a shelled rhizopod belonging to the group Imperforata, family Arcellina, with the following results, to two decimal places :—

	Mean.	S.D.	C. of V.
Outer diameter .	55·79	5·73	10·27 per cent.
Inner „ .	15·91	2·17	13·66 „

Thus, judging by the S.D. column, giving the absolute size of deviation, the outer diameter would appear to be more variable than the inner, but the C. of V. column shows that, if we take the sizes of the two diameters into account, the inner is really the more variable of the two. To turn aside the edge of possible criticism it should be added that the authors also give the errors to which the above measures are subject, as unless these are known we cannot tell whether the differences observed in variation are significant or not of a real difference in fact, but that question must be left until the theory of errors due to sampling has been developed in a later chapter.

The C. of V. varies considerably for different characters. W. R. Macdonell states that '3 to 5·5 are representative values for variability in man, while in plants it may run to 40,' and Pearson and others have shown that for stature in man it varies from about 3 to 4 and for the length of long bones from 4 to 6.

CHAPTER VII

FREQUENCY DISTRIBUTION: EXAMPLES TO ILLUSTRATE CALCULATING AND PLOTTING: SKEWNESS

Calculation of Mean and Standard Deviation. *Example* (1).—We return now to the examination example in order to show how the labour of calculation in finding the arithmetic mean and standard deviation of a frequency distribution may be somewhat lessened.

The various steps in the process appear in Table (14). In the first column the marks at the middle of each class-interval have been written down, and we make the assumption that all the candidates in any one class have the same number of marks, namely, the marks at the middle of the class-interval. In any case where the number of observations is large, and where the class-intervals are reasonably small, the errors resulting from such an assumption will be insignificant, because the individuals in each class are just as likely to have values above as below the value at the middle of the class-interval, and they will therefore compensate for one another.

We now seek to alter the scale of marking so as to produce a simpler set of marks than the original, which will make the work of finding the mean also simpler, but we must not forget at the end to change back again to the original scale. We choose a number from col. (1), somewhere near the required mean, to act as a kind of origin from which to measure the other numbers in the column. This choice is only a rough guess, and it is really immaterial which number is selected as origin, except that the nearer it is to the mean the lighter will be the calculation to follow; the number 33 has been selected in this instance.

In col. (2) are written down the deviations of the marks in each class from 33, so that now some candidates appear as if they were 5, 10, 15 . . . marks to the bad, and others as if they were 5, 10, 15 . . . to the good. So long as we remember to add 33 at the end we can content ourselves therefore by finding the mean of the marks as given in col. (2). But these again can be further simplified by dividing each candidate's marks by 5, and we then only need

to find the mean of the marks as shown in col. (3), so long as we remember to multiply by 5 at the first step back to the old scale of marking. The addition of col. (5) makes it easy to calculate this mean, for it gives the result of multiplying each value of the variable (the number of marks in each class) by its appropriate weight (the number of candidates who obtained that number of marks).

TABLE (14). MARKS OBTAINED BY 514 CANDIDATES IN A CERTAIN EXAMINATION—(ANALYSIS OF METHOD FOR CALCULATING MEAN AND STANDARD DEVIATION).

(1) Marks on old scale.	(2) Deviation of Nos. in Col. (1) from 33.	(3) Marks on new scale.	(4) Frequency of Candidates.	(5) Product of Nos. in Cols. (3) & (4).	(6) Product of Nos. in Cols. (3) & (5).
		(x)	(f)	(fx)	(fx^2)
3=33−30	−30	−6	5	− 30	180
8=33−25	−25	−5	9	− 45	225
13=33−20	−20	−4	28	−112	448
18=33−15	−15	−3	49	−147	441
23=33−10	−10	−2	58	−116	232
28=33− 5	− 5	−1	82	− 82	82
33=33	..	..	87	..	..
38=33+ 5	+ 5	+1	79	+ 79	79
43=33+10	+10	+2	50	+100	200
48=33+15	+15	+3	37	+111	333
53=33+20	+20	+4	21	+ 84	336
58=33+25	+25	+5	6	+ 30	150
63=33+30	+30	+6	3	+ 18	108
..	..	..	514	−110	2814

Thus, on this new scale, the mean marks obtained are

$$=\frac{5(-6)+9(-5)+28(-4)+\ \ldots\ +87(0)+\ \ldots\ +6(+5)+3(+6)}{514}$$

$$=\frac{-532+422}{514}$$

$$=\frac{-110}{514}$$

$$=-0{\cdot}214.$$

This, then, is the mean of the marks obtained by the candidates on the scale indicated in col. (3). If the marks are on the scale given in col. (2), the mean is 5(−0·214), *i.e.* −1·070. To bring them back to the original scale as in col. (1) we must add 33 to this result, so that the required arithmetic mean

$$\begin{aligned}&=33+5(-0{\cdot}214)\\&=33-1{\cdot}070\\&=31{\cdot}93.\end{aligned}$$

To find the Standard Deviation, or the root-mean-square deviation from the arithmetic mean, it is convenient as before to work with the simplified scale, to measure the deviations from the arbitrary origin (33) associated with that scale, and to make the necessary corrections at the end of the work.

Col. (5) in Table (14) gives the deviation multiplied by the frequency in each class, the frequency denoting the number of times the particular deviation occurs. Hence, if these numbers be multiplied again by the numbers in col. (3), we shall have each separate deviation squared and multiplied by its frequency. The results are shown in col. (6), and they must be added, and their sum divided by the sum of the frequencies (514), to give the mean-square deviation, which we may represent by s^2.

Thus $$\begin{aligned}s^2&=2814/514\\&=5{\cdot}475,\end{aligned}$$

and this is the mean-square deviation referred to 33 as origin. We require the corresponding expression referred to the mean, 31·93, as origin. If we denote this by s_m^2 there is a simple relation connecting the two, namely,

$$s_m^2=s^2-\bar{x}^2,$$

where $\bar{x}$ is the deviation of the mean itself from 33 [see Appendix, Note 5]; of course s_m, s, and $\bar{x}$ are all to be measured on the same scale, the simplified scale adopted with 5 marks as unit.

Now we have already shown that the deviation of the mean from 33 = −0·214, and this is therefore the value of $\bar{x}$.

Hence $$\begin{aligned}s_m^2&=5{\cdot}475-(-0{\cdot}214)^2\\&=5{\cdot}475-0{\cdot}046\\&=5{\cdot}429\\&=(2{\cdot}33)^2.\end{aligned}$$

And, returning to the old scale, the standard deviation, usually denoted by σ

$$=5(2{\cdot}33)$$
$$=11{\cdot}65.$$

We notice that $3\sigma=34{\cdot}95$, and this range on either side of the mean amply takes in all the observations.

The mean deviation is readily found from Table (14) by adding up the numbers in col. (5) regardless of sign and dividing by the sum of frequencies, 514.

Thus, on the new scale, the mean deviation

$$=\tfrac{954}{514}$$
$$=1{\cdot}856,$$

which, on the old scale, becomes 5(1·856) or 9·28. This, however, is the mean deviation measured from 33 as origin, and a correction has to be applied to get the mean deviation measured from the median or from the mean.

To get the mean deviation from the mean we note that the difference between the mean, 31·93, and 33 is 1·07. Hence it should be clear from Table (14) that, by measuring from 33 instead of from 31·93, we have made the deviations of all the marks from 33 upwards too little by 1·07, and we have made the deviations of all the marks from 28 downwards too much by 1·07. Hence, to get the deviation required we must add to 9·28 an amount

$$=\tfrac{1}{514}[1{\cdot}07(87+79+\ .\ .\ .\ +3)-1{\cdot}07(82+58+\ .\ .\ .\ +5)]$$
$$=\frac{1{\cdot}07}{514}(283-231)$$
$$=\frac{1{\cdot}07}{514}\times 52$$
$$=0{\cdot}108.$$

Therefore, the mean deviation measured from the mean $=9{\cdot}39$. This may be compared with $\tfrac{4}{5}$(standard deviation)$=9{\cdot}32$.

Also the quartile deviation for this distribution has been shown to be$=8{\cdot}02$, and it may be compared with $\tfrac{2}{3}$(standard deviation) $=7{\cdot}77$.

Plotting of a Frequency Distribution. The data for the two examples which follow are taken from the *Quarterly Return of Marriages, Births, and Deaths*, No. 261, issued by the Registrar-General.

The first shows the proportion to population of cases of infectious disease notified in 241 large towns of England and Wales for the thirteen weeks ended 4th April 1914. This proportion was given for each town separately in the Return, but, in order to bring out the distinctive features of the distribution, the several towns have

TABLE (15). PROPORTION TO POPULATION OF CASES OF INFECTIOUS DISEASE NOTIFIED IN 241 LARGE TOWNS OF ENGLAND AND WALES DURING THE THIRTEEN WEEKS ENDED 4TH APRIL 1914.

<table>
<tr><th>Case Rate per 1000 persons living.</th><th>Each dot below represents One Town with Notified Rate of Infectious Disease between limits as given in previous column.</th><th>Total No. of Towns with given Rate.</th></tr>
<tr><td>0—</td><td>....|</td><td>5</td></tr>
<tr><td>2—</td><td>....|....|....|....|....|....|....|....</td><td>39</td></tr>
<tr><td>4—</td><td>....|....|....|....|....|....|....|....|....|....|....|....|....|....</td><td>69</td></tr>
<tr><td>6—</td><td>....|....|....|....|....|....|....|....|.</td><td>41</td></tr>
<tr><td>8—</td><td>....|....|....|....|....|....</td><td>29</td></tr>
<tr><td>10—</td><td>....|....|....|....|..</td><td>22</td></tr>
<tr><td>12—</td><td>....|....|....|.</td><td>16</td></tr>
<tr><td>14—</td><td>....|..</td><td>7</td></tr>
<tr><td>16—</td><td>....|</td><td>5</td></tr>
<tr><td>18—</td><td>...</td><td>3</td></tr>
<tr><td>20—</td><td>....</td><td>4</td></tr>
<tr><td>22—</td><td></td><td>0</td></tr>
<tr><td>24—</td><td></td><td>0</td></tr>
<tr><td>26—</td><td>.</td><td>1</td></tr>
<tr><td></td><td></td><td>241</td></tr>
</table>

been, in Table (15), represented by dots and put into different classes *according to the proportion* of infectious cases *notified in each*, with a separate line for each class: *e.g.* if the proportion for any town was 5·37 a dot was placed in the line corresponding to the class of towns for which the rate was '4 and less than 6.' Every

fifth dot in each line was ticked off, so as to make them easy to count up and also to keep the lines, down the paper as well as across, straight. The frequency, *i.e.* the number of dots in each class, was then recorded in a column at the extreme right-hand side of the paper.

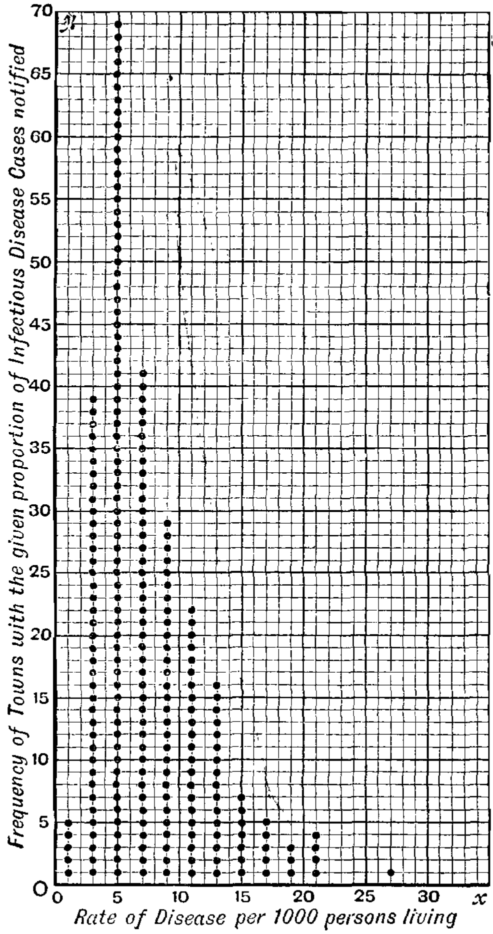


Fig. (1).

It will be at once seen that this procedure, without calculating any averages, etc., ultimately gives to the eye a very good picture of the distribution, and indeed it is the basis of the graphical method of studying statistics. In drawing a proper graph we use a specially ruled sheet of paper which is divided up into a large number of equal small squares by 'horizontal' (cross) and 'vertical' (up-and-

down) lines. This merely enables us to place our dots accurately in position, as shown in fig. (1), where the numbers 0, 5, 10 . . . have been marked off along the line $0x$ to correspond to 'case

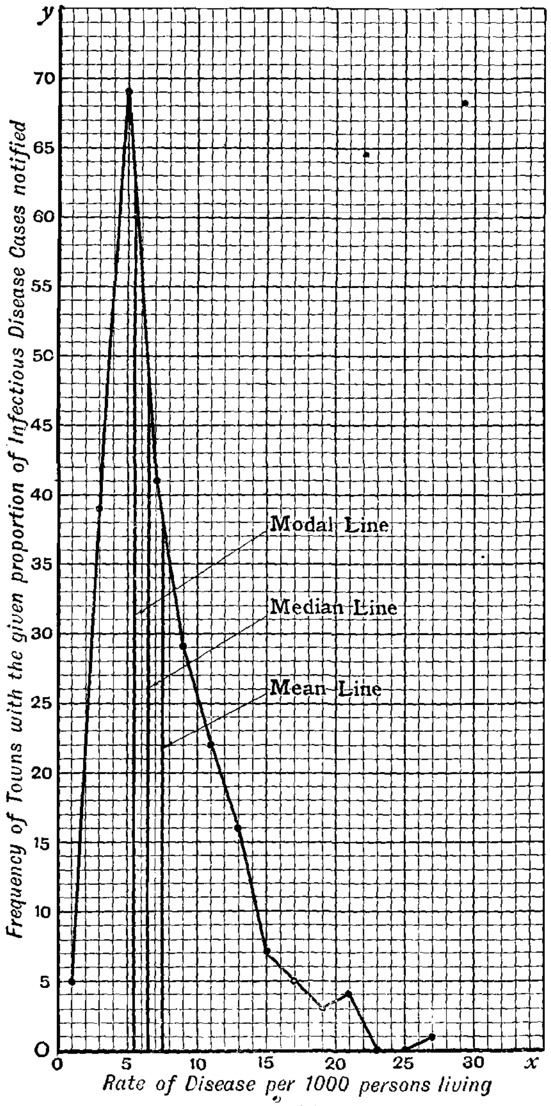


FIG. (2).

rates' of these magnitudes: thus rates of '4 and less than 6' were recorded by 69 successive dots along a vertical line at a distance 5 (the centre of the class-interval 4-6) from the axis $0y$.

The final configuration in fig. (1), when turned half round, is exactly the same as that of Table (15). If desired the frequency

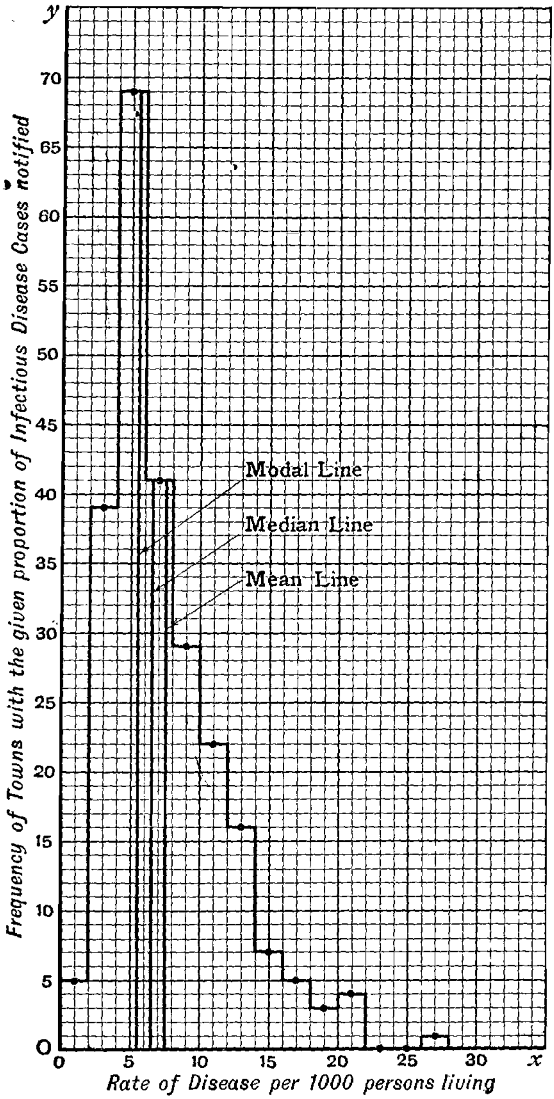


FIG. (8).

may be recorded, dot by dot, on a side piece of paper and then only the topmost dot in each class need be marked on the graph sheet. In order, however, to enable the eye to measure the height

of each frequency in relation to the rest, it is advisable in that case to connect up adjacent dots as in fig. (2) or as in fig. (3).

The last method of representation (fig. (3)), to which the name ***histogram*** has been given by Professor Karl Pearson, is particularly useful and should be carefully studied. It is formed in this case by erecting a succession of rectangles with the lines 02, 24, 46 . . . along Ox as their bases, corresponding to the successive classes of the given distribution, and with heights proportional to the frequencies proper to those classes. It is not necessary to complete the sides of the rectangles, but, if they were completed, each would enclose a number of squares proportional to the frequency of towns with the rate of disease defined by its base: *e.g.* the first rectangle would enclose 10 squares, the second 78, the third 138, and so on, numbers respectively proportional to 5, 39, 69, and so on. It follows that the total area enclosed between the histogram and the axis Ox is proportional to the aggregate frequency of towns observed.

Now we might conceive a step further taken and a smoothed curve drawn freehand so as to agree as closely as possible with fig. (2) or fig. (3), but with all the sharp corners smoothed out, and so nicely adjusted as to make the area enclosed between the curve, the axis Ox, and lines parallel to Oy defining the limits of any class, proportional to the frequency of towns in that class. To this fig. (2) and fig. (3) might be regarded as approximating if only a sufficient number of observations were recorded, and only in that case would it be possible to draw it with any accuracy. Such a curve is called a *frequency curve*, measuring as it does the frequency of the observations in different classes.

[Assuming that corresponding to a given frequency distribution a curve of this kind does really exist—and the assumption turns upon the frequency being continuous—the reader who is acquainted with the notation of the Calculus will recognise that, if (x, y) represents any point on the curve, $y\delta x$ measures the frequency of observations or measurements of an organ or character lying between the values x and $(x+\delta x)$, when the total frequency comprises a large number of observations, say 500 to 1000.

Further, it will appear later that the mean, the median, and the mode have a geometrical interpretation of no small importance associated with the curve.

The mean x corresponds to the particular ordinate y which passes through the centroid or centre of gravity of the area between the frequency curve and axis Ox, because

$$\text{the mean} = \underset{\delta x \to 0}{L^t} \Sigma(x \,.\, y\delta x) \Big/ \underset{\delta x \to 0}{L^t} \Sigma(y\delta x),$$

where the summation extends throughout the distribution,

$$= \int xydx \Big/ \int ydx$$

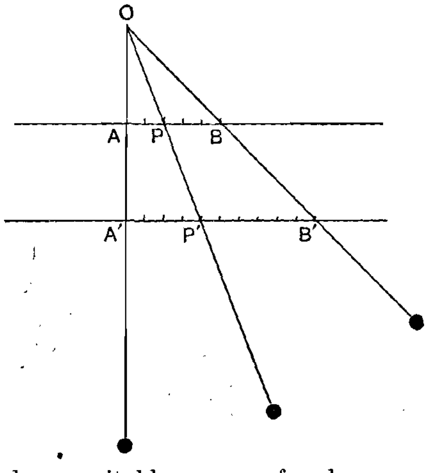
where the integral extends throughout the curve.

The median x corresponds to the ordinate y which bisects this same area; *e.g.* in fig. (3), the number of small squares on either side of the median in the space bounded by the histogram and the axis represents half the total number of observations, two small squares corresponding to each observation.

The mode x corresponds to the maximum ordinate of the curve, measuring the greatest frequency in the whole distribution.]

Skewness. There is one feature of a frequency distribution which catches the eye sooner almost than any other, and that is its symmetry or lack of symmetry. It is important therefore that we should have some means of measuring it.

In a symmetrical distribution the mean, mode, and median coincide, and we have, as it were, a perfect balance between the frequency of observations on either side of the mode or ordinate of maximum frequency. In a skew distribution the centre of gravity is displaced and the balance thrown to one side: the amount of this displacement measures the skewness. But there is another factor to be taken into account, for when the variability of the distribution is great the balance is more sensitive than when it is small, and the difference between mean and mode is consequently more pronounced though it may not be significant of any greater skewness. This will be clear in the light of the analogy of the swing of a pendulum. If OPP′ denote the pendulum in the accompanying figure, OAA′ its mean position, and OBB′ an extreme position, the displacement in the position OPP′ from the mean, if measured along the scale AB, is AP, and, if measured along the scale A′B′, is A′P′. But, since the amount of swing in either case is the same, it would be more appropriate to write the linear displacement as a fraction of the full swing so as to make these two measures also the same, thus



$$AP/AB = A'P'/A'B'.$$

So, in the case of a frequency distribution, Professor Karl Pearson has suggested as a suitable measure for skewness, not the difference between mean and mode, but the ratio of this difference to the variability. Thus

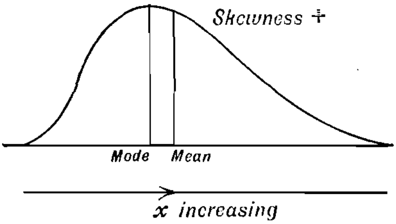
$$skewness = (mean - mode)/S.D.$$

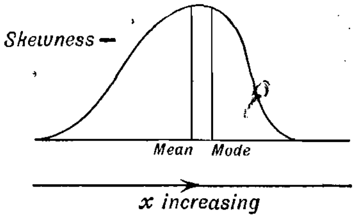
or, approximately,

$$=3(\text{mean}-\text{median})/\text{S.D.} \text{ (see p. 39)},$$

a form which is sometimes useful.

According to this convention the skewness is regarded as positive





when the mean is greater than the mode, and as negative when the mode is greater than the mean.

Illustrations of frequency curves, with the position of mode and mean marked, will be found in Chapter XVII.

We proceed to the detailed calculations necessary in the infectious diseases example.

TABLE (16). PROPORTION TO POPULATION OF CASES OF INFECTIOUS DISEASE NOTIFIED IN 241 LARGE TOWNS OF ENGLAND AND WALES DURING THE THIRTEEN WEEKS ENDED 4TH APRIL 1914.

(1) Case Rate per 1000 persons living.	(2) Deviation from 7.	(3) Frequency of Towns with given Rate.	(4) Product of Nos. in Cols. (2) & (3).	(5) Product of Nos. in Cols. (2) & (4).
	(x)	(f)	(fx)	(fx^2)
0 and less than 2	−3	5	−15	45
2 ,, ,, 4	−2	39	−78	156
4 ,, ,, 6	−1	69	−69	69
6 ,, ,, 8	..	41	..	..
8 ,, ,, 10	+1	29	+29	29
10 ,, ,, 12	+2	22	+44	88
12 ,, ,, 14	+3	16	+48	144
14 ,, ,, 16	+4	7	+28	112
16 ,, ,, 18	+5	5	+25	125
18 ,, ,, 20	+6	3	+18	108
20 ,, ,, 22	+7	4	+28	196
26 ,, ,, 28	+10	1	+10	100
..	..	241	+68	1172

Example (2).—The various averages and measures of variability of the distribution can be calculated just as in the case of the last example, and the data required to determine the mean and the standard deviation are set out in Table (16). We can afford now to miss out some of the more obvious steps in explanation.

On the scale of col. (2), where a difference of 2 in the case rate, per 1000 persons living, is the unit and where a case rate of 7 is taken as origin, the *mean*, by the result of col. (4)

$$=\tfrac{68}{241}$$
$$=0{\cdot}282.$$

Hence, on the original scale, the mean

$$=7+2(0{\cdot}282)$$
$$=\mathbf{7{\cdot}564}.$$

Again, the mean-square deviation, on the scale of col. (2), measured from 7 as origin is

$$s^2=\tfrac{1172}{241}$$
$$=4{\cdot}863\;;$$

and $\bar{x}$, the deviation of the mean from 7 as origin, on the scale of col. (2)=0·282. Thus the mean-square deviation measured from the mean,

$$s_m{}^2=s^2-\bar{x}^2$$
$$=4{\cdot}863-(0{\cdot}282)^2$$
$$=4{\cdot}783.$$

Therefore, the *standard deviation* σ, on the original scale

$$=2\sqrt{4{\cdot}783}$$
$$=\mathbf{4{\cdot}374}.$$

Since $3\sigma=13{\cdot}122$, the range '(mean-3σ) to (mean$+3\sigma$)' includes all but one or two observations.

To determine the median, we conceive the towns ranged in order according to the proportion of infectious cases notified in each, from the least to the greatest, and the town with the median rate is the 121st from either end.

But the 113th town has a notified case rate of approximately 6 per 1000, and the 154th town has a notified case rate of approximately 8 per 1000.

Thus a difference of 41 towns corresponds to a difference of 2 in the rate, hence a difference of 8 towns corresponds to a difference of 0·39 in the rate; therefore the *median rate*=**6·39** *approximately.*

By referring to the original records and writing down the rate

for each town in the group ' rate 6 and less than 8 ' in which the median lay, the accurate value of the median turned out to be 6·30.

The lower quartile or case rate of the imaginary town, No. $\frac{1}{4}(241)$, or 60·25, one-quarter way along the ordered sequence of towns, is readily shown to be **4·47**, and the *upper quartile* or case rate of *town No.* $\frac{3}{4}$*(241), or 180·75, is* ***9·84***.

Hence the *quartile deviation*

$$=\tfrac{1}{2}(9{\cdot}84-4{\cdot}47)$$
$$=\mathbf{2{\cdot}69}.$$

With this may be compared $\frac{2}{3}$(S.D.)$=\frac{2}{3}(4{\cdot}37)=2{\cdot}92$.

Again, the *mean deviation* measured from 7

$$=2(\tfrac{392}{241})$$
$$=3{\cdot}253.$$

Measured from the mean, it becomes

$$=3{\cdot}253+\frac{0{\cdot}564}{241}[(41+69+39+5)-(29+22+16+7+5+3+4+1)]$$
$$=3{\cdot}253+(0{\cdot}564)(67)/241$$
$$=\mathbf{3{\cdot}41}$$

and this may be compared with $\frac{4}{5}$(S.D.)$=\frac{4}{5}(4{\cdot}374)=3{\cdot}50$.

If we estimate the *mode* by inspection of the frequency graphs in figs. (2) and (3), we should say it comes between 5 and 6 ; supposing we call it 5·5, very roughly.

In this case, taking the values actually calculated for mean and median,

$$(\textit{mean}-\textit{mode})=7{\cdot}56-5{\cdot}50$$
$$=2{\cdot}06,$$

and

$$3(\text{mean}-\text{median})=3(7{\cdot}56-6{\cdot}39)$$
$$=3(1{\cdot}17)$$
$$=3{\cdot}51\ ;$$

so that the rule

$$(\text{mean}-\text{mode})=3(\text{mean}-\text{median})$$

is far from being true according to these results ; this is partly due, of course, to the very unsymmetrical character of the distribution.

The relative positions of the mean, median, and modal points as calculated are indicated in figs. (2) and (3) by three lines drawn parallel to Oy through these points to meet the graph.

Finally, *skewness*$=$(mean$-$mode)/S.D.$=2{\cdot}06/4{\cdot}37=0{\cdot}47$.

Example 3.—The next example deals with the deaths of infants under one year, out of every thousand born, in 100 great towns in the United Kingdom during the thirteen weeks ended 4th April 1914.

The details of the calculation may be left in this case to the reader, who is recommended to follow the method shown in the last example so far as possible throughout, including the plotting of the distribution in different ways. The statistics are as follows :—

TABLE (17). DEATH RATE OF INFANTS UNDER 1 YEAR PER 1000 BIRTHS.

(1) Death Rate.	(2) No. of Towns with Death Rate as in Col. (1).	(3) Death Rate.	(4) No. of Towns with Death Rate as in Col. (3).
30 and under 40	1	120 and under 130	16
50 „ 60	3	130 „ 140	11
60 „ 70	2	140 „ 150	10
70 „ 80	6	150 „ 160	8
80 „ 90	7	160 „ 170	3
90 „ 100	6	170 „ 180	1
100 „ 110	11	200 „ 210	1
110 „ 120	13	240 „ 250	1

The more important results are :—
Arithmetic mean=118·9 ; S.D.=32·2 ;
median=120·9 ; quartile deviation=19·5.

Example (4).—As another example corresponding details may be worked out for the following temperature records taken at noon at a certain spot in Chester week by week during a period of time covering five years, the results in this case being :—

mean=55·10 ; S.D.=10·33 ;
median=54·88 ; quartile deviation=7·94

TABLE (18). 257 WEEKLY RECORDS OF TEMPERATURE (FAHRENHEIT).

(1) Temperature Limits in Degrees.	(2) No. of Records between Limits shown in Col. (1)	(3) Temperature Limits in Degrees.	(4) No. of Records between Limits shown in Col. (3)
25·5–29·5	1	53·5–57·5	30·5
29·5–33·5	1	57·5–61·5	31·5
33·5–37·5	9	61·5–65·5	30
37·5–41·5	11·5	65·5–69·5	26
41·5–45·5	28	69·5–73·5	13·5
45·5–49·5	31·5	73·5–77·5	4
49·5–53·5	36·5	77·5–81·5	3

Before closing the chapter a slightly different manner of graphing the statistics is worth noticing, as it provides us with a fairly quick though rough alternative method of determining the mode and median.

Take, for example, the examination marks data which for this purpose must first be thrown into the second form shown below Table (7). We mark off on some convenient scale along OX distances 5, 10, 15, 20 . . . 65 from O to represent these numbers of marks respectively, and at the points obtained we erect lines parallel to OY of lengths 5, 14, 42, 91 . . . 514 to represent the numbers of candidates who obtained not more than 5, 10, 15, 20 . . . 65 marks respectively. A freehand curve is then drawn through the summits of these lines in the manner indicated in fig. (4), starting from a height 5 and rising to a height 514 above the axis OX. It is called an ogive curve.

By means of this curve we can approximately state at once how many candidates obtained any given number of marks *or less*. Suppose, for example, we wish to know how many candidates obtained 22 marks or less, we have only to measure off a distance 22 from O, represented by ON, and erect a perpendicular NP to meet the curve at P. Since NP=110 we infer from the manner in which the curve has been formed that 110 candidates obtained 22 marks or less, so that, incidentally, the 110th candidate from the bottom must have obtained approximately 22 marks. This suggests that by working backwards we can also read off roughly the number of marks gained by any particular candidate when his order in the list is known. Thus, to find the median, *i.e.* the marks due to candidate No. 257·5, we merely draw a line parallel to OX at a height 257·5 above it and the portion of this line cut off between the curve and OY measures the median. The value given by this method is approximately 31·5. Similarly the quartiles are found by drawing lines parallel to OX at heights 128·5 and 385·5 above it with results about 23·3 and 39·2 respectively.

Again, as we gradually increase the number of marks, the number of candidates getting that number of marks or less must increase also, but the rate of this second increase is variable. The reader will perceive that where the height above OX changes slowly the gradient of the curve is small, but where it changes by big steps the gradient is steep, and it is at its steepest just in the neighbourhood where the greatest addition is being made to the height as the marks increase, *i.e.* where the frequency of additional candidates is at its greatest, so determining the mode: this should be

clear on a comparison of the two arrangements of the data in and below Table (7). By sliding a straight-edge along the contour of the curve we can estimate approximately where the curve is steepest, for at this point the direction of turning of the ruler or

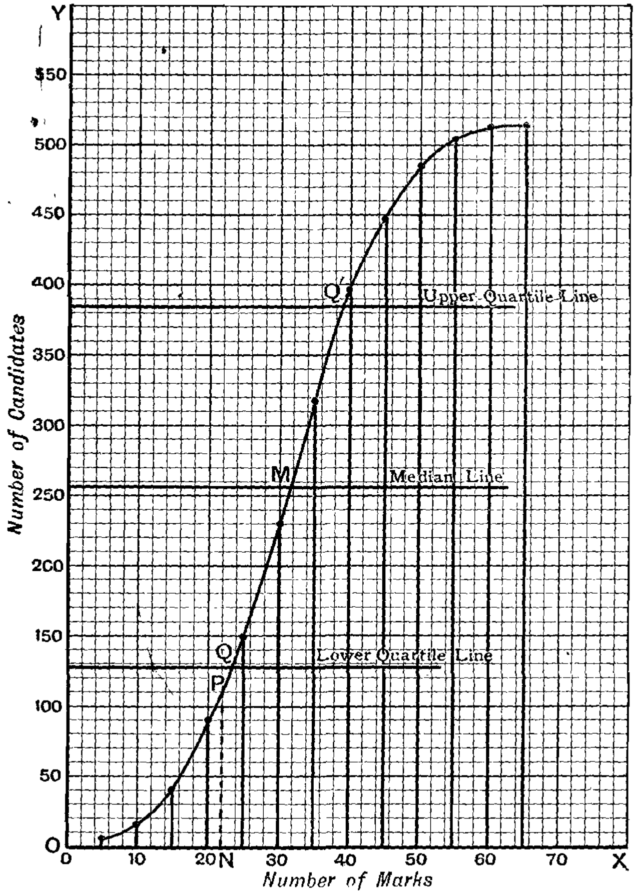


FIG. (4). Graph showing the Number of Candidates who obtained not more than any given Number of Marks.

straight-edge must change. This gives for the mode a value in the neighbourhood of 32.

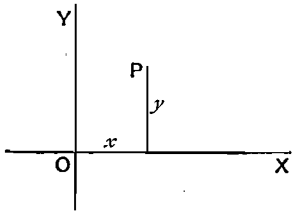
It might be advisable to treat the other examples by this method also, so as to compare results.

CHAPTER VIII

GRAPHS

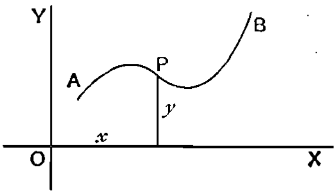
From the mathematical point of view graphs may be regarded as the alphabet of Algebraical Geometry.

We can locate a point in a plane, relative to two perpendicular lines or axes as they are called, OX, OY, which serve as boundaries



of measurement, when we know y and x, its shortest distances from these boundaries. This fact serves to connect up Geometry, in which points are elements, with Algebra, in which x's and y's, standing always for numbers, are elements. The names *abscissa* (*ab*—from, and *scindo*—I cut) and *ordinate* are given to x and y, or, when we refer to them together, they may be spoken of as the *co-ordinates* of P.

The celebrated French philosopher, Descartes (1596-1650), was the founder of Cartesian Geometry, and if we may venture to compress the essence of his system into a single statement, it is this—When a point P is free to take up *any* position in a given plane, its x and y are quite independent: they may be allotted any values irrespective of one another. Suppose, however, that P is constrained to lie somewhere on an assigned curve, such as APB in the figure, then x and y are no longer independent, for, so soon as x is fixed, y is fixed also; it follows that in this case some relation, algebraical or otherwise, such as $y=x^3-2x+7$, must exist between x and y, and the relation may be called the equation of the curve which gives rise to it.

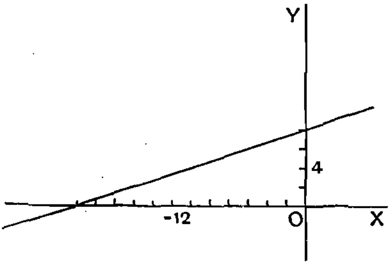


Now, if to every curve there corresponds in this way some equation and to every equation some curve, it seems likely that the simpler the curve the simpler will be the corresponding equation, and *vice versa*. In fact, the student who does not know it already

need only refer to the most elementary treatise on graphs to find that every equation of the first degree in x and y, *i.e.* one which does not involve any x^2, y^2, xy, or higher powers, represents some straight line. Any such equation, *e.g.*

$$x-3y+12=0,$$

can be at once thrown into either the form



(1) $$\frac{x}{-12}+\frac{y}{4}=1,$$

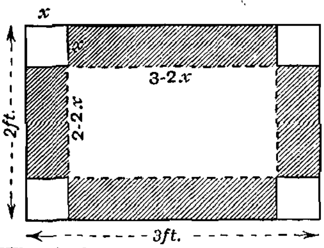
where -12 and 4 are intercepts made by the line on the axes OX and OY ; or

(2) $$y=\tfrac{1}{3}x+4,$$

where $\frac{1}{3}$, *i.e.* 1 in 3, is the measure of its gradient and 4 the height above the origin at which it cuts the axis OY.

Further, every equation of the second degree in x and y, which may involve x^2, y^2, and xy, but no higher powers, represents geometrically some conic, a family of curves comprising the parabola, the ellipse, and the hyperbola, with the circle and two straight lines as particular cases. The earth and other planets, likewise comets, in their journeys through space travel along curves belonging to the same family, one of ancient and historical connections.

These conics need not, however, detain us, and we pass on at once to an example of a cubic graph to show how a very little knowledge of the theory may be put to some practical use. Suppose a box manufacturer has a large number of rectangular sheets of cardboard, 3 ft. long by 2 ft. broad, and he wishes to make open boxes with them by cutting a square piece of the same size out of each corner and turning up the flaps that are left. How big should the squares be if this is to be done with as little waste as possible ? Clearly this is commercially an important type of problem to solve.



[The shaded flaps are bent upwards along the dotted lines.]

Let us denote a side of the square to be cut out of each corner by x feet. Then the bottom of the required box will have dimensions

$$(3-2x) \text{ ft. by } (2-2x) \text{ ft.}$$

and its depth will be x ft.

Hence the capacity of the box when completed will be

$$x(3-2x)(2-2x) \text{ cu. ft.},$$

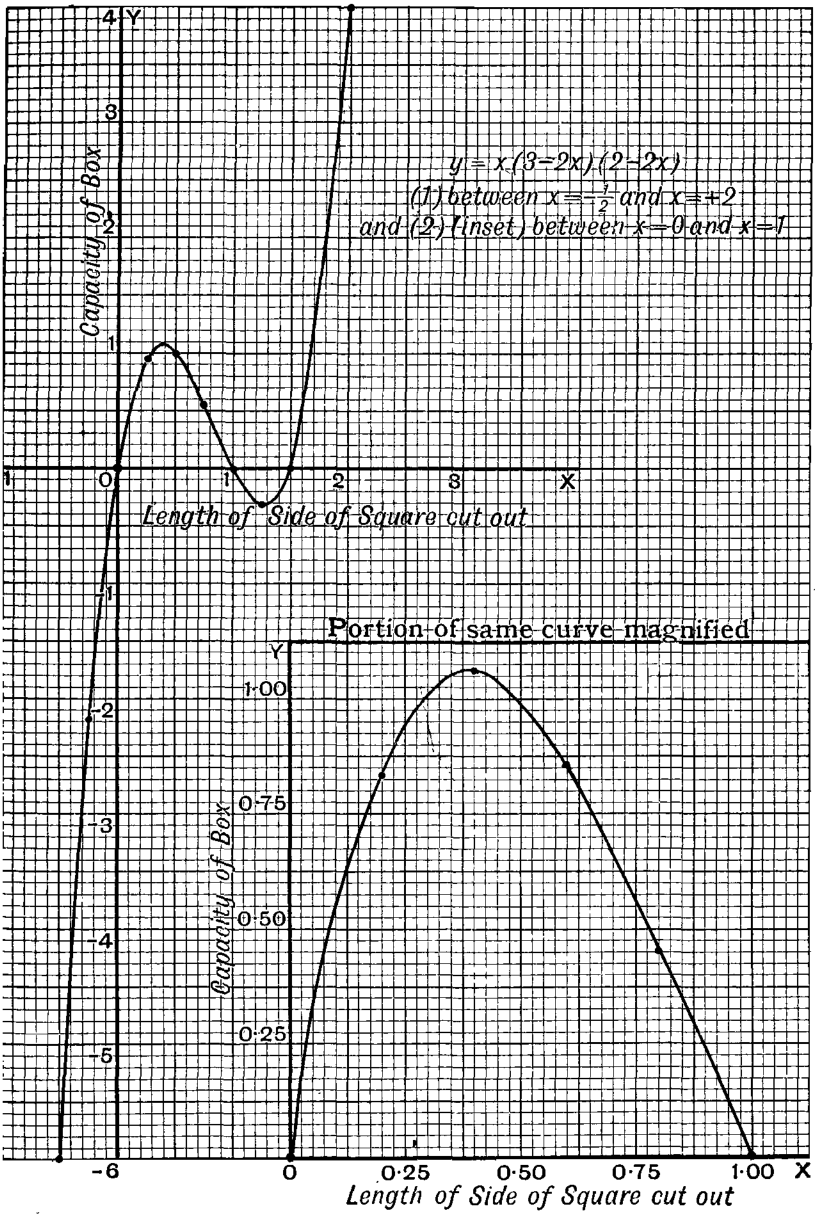
and he makes best use of the material who produces the most capacious box. Call this expression y and let us find the values of y corresponding to different values of x so as to be able to draw roughly the curve of which the equation is

$$y=x(3-2x)(2-2x) \quad . \quad . \quad . \quad (1)$$

TABLE (19). TABLE OF CORRESPONDING VALUES OF x AND y IN THE CURVE $y=x(3-2x)(2-2x)$.

x	$2x$	$(3-2x)$	$(2-2x)$	$x(3-2x)(2-2x)$	y
-1	-2	5	4	-20	-20
$-\frac{1}{2}$	-1	4	3	-6	-6
$-\frac{1}{4}$	$-\frac{1}{2}$	$\frac{7}{2}$	$\frac{5}{2}$	$-\frac{35}{16}$	$-2{\cdot}19$
0	0	3	2	0	0
$+\frac{1}{4}$	$+\frac{1}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$+\frac{15}{16}$	$+0{\cdot}94$
$+\frac{1}{2}$	$+1$	2	1	$+1$	$+1$
$+\frac{3}{4}$	$+\frac{3}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	$+\frac{9}{16}$	$+0{\cdot}56$
$+1$	$+2$	1	0	0	0
$+1\frac{1}{4}$	$+\frac{5}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{5}{16}$	$-0{\cdot}31$
$+1\frac{1}{2}$	$+3$	0	-1	0	0
$+2$	$+4$	-1	-2	$+4$	$+4$
$+2\frac{1}{2}$	$+5$	-2	-3	$+15$	$+15$
0·2	0·4	2·6	1·6	(0·2)(2·6)(1·6)	0·83
0·4	0·8	2·2	1·2	(0·4)(2·2)(1·2)	1·06
0·6	1·2	1·8	0·8	(0·6)(1·8)(0·8)	0·86
0·8	1·6	1·4	0·4	(0·8)(1·4)(0·4)	0·45
0·38	0·76	2·24	1·24	(0·38)(2·24)(1·24)	1·055
0·39	0·78	2·22	1·22	(0·39)(2·22)(1·22)	1·056
0·40	0·80	2·20	1·20	(0·40)(2·20)(1·20)	1·056
0·41	0·82	2·18	1·18	(0·41)(2·18)(1·18)	1·055

We get a tolerably good idea of the shape of the curve by plotting the points (x, y) shown in Table (19) from $x=-\frac{1}{2}$ to $x=+2$ as in fig. (5). It is simply a matter of practice to be able to determine the whole curve from a few points in this way, and the greater the number of points plotted the more accurately will it be possible to draw the curve. It should be noticed that the points for which $y=0$ are in a sense key-points to the curve: they are readily


$y = x(3-2x)(2-2x)$
(1) between $x = -\frac{1}{2}$ and $x = +2$
and (2) (inset) between $x = 0$ and $x = 1$
Capacity of Box
Length of Side of Square cut out
Y
X
4
3
2
1
O
-1
-2
-3
-4
-5
-6
1
2
3
Portion of same curve magnified
1·00
0·75
0·50
0·25
0
0·25
0·50
0·75
1·00
Capacity of Box
Length of Side of Square cut out

found by making the factors separately zero in the right-hand side of equation (1), namely $x=0$, $3-2x=0$, and $2-2x=0$, and by plotting them first they serve as a guide to the position of points subsequently plotted.

We want to know for what value of x the capacity of the box, y, is greatest and the preliminary plotting is enough to indicate a maximum value for y between $x=0$ and $x=1$, for the curve first rises and then falls between these two limits. In order to discover more exactly where the maximum is located we therefore plot in addition the points corresponding to $x=0{\cdot}2$, $0{\cdot}4$, $0{\cdot}6$, $0{\cdot}8$ respectively, and this is done on a larger scale than that used in the first diagram because the accuracy is thereby increased (see fig. (5) inset).

The calculations and figure suggest that the maximum required is very near the point for which $x=0{\cdot}4$, so we next work out values of y in this neighbourhood, corresponding, say, to $x=0{\cdot}38$, $0{\cdot}39$, $0{\cdot}40$, $0{\cdot}41$, with the results shown at the foot of Table (19). From these we conclude that to a fair degree of accuracy the maximum value of y is given by taking $x=0{\cdot}395$. It would be possible in the same way to calculate more decimal places, but we have gone far enough to make the method clear.

Hence the side of each square cut out should be of length

$$0{\cdot}395 \text{ ft., or } 4\tfrac{3}{4} \text{ in.}$$

Whenever the value of one variable, y, depends upon that of another variable, x, in such a way that when x is given y is known, so that y may be termed a function of x, corresponding values of x and y can be plotted—as was done in the example just discussed—and a curve drawn by joining up the points obtained, the relation which connects x and y being the equation of this curve. Moreover, it is possible, by calculating enough points from the equation and plotting them, to get the curve as accurately as we please.

In Statistics, however, we usually have to start the other way round and reach the equation, if at all, last. We make observations of two sets of variables, a set of x's, and a set of y's, one of which is dependent in some way upon the other—*e.g.* y, the dependent variable, might denote the number of individuals observed to have a certain organ of length x, the independent variable—and thus we get pairs of corresponding values like (x_1, y_1), (x_2, y_2), (x_3, y_3) . . . We met with examples of this method of recording results in the last chapter, and we need only repeat here that its chief virtue is suggested in the root of the word itself—it is more *graphic* than a

long table of figures and, by means of it, many of the essential features of a problem are immediately seized upon.

Now for some purposes it may be necessary to go further and to find what curve would best fit the points plotted, assuming they were numerous enough, and what equation between x and y would best describe the curve. But the graphs we meet in Statistics, bearing, for instance, upon sociological or biological problems, are in general much more wayward than the mathematical kind we have referred to in the present chapter: it is impossible to set down simple equations to which they can be rigidly confined, and when we are unable to find any relation which accurately and uniquely defines y as a function of x we must rest satisfied with the most manageable equation and the best fit we can get.

In sciences such as Engineering and Physics it is often possible to fix upon two mutually dependent variables, x and y, and to observe enough corresponding values of each to enable us to draw a graph which answers very closely to the true relationship between them, so that a connecting equation can be determined; *e.g.* we may plot the amount of elastic stretch, y, in a wire when different weights, x, are hung from the end of it, and it is found that y is directly proportional to x. If we deal in this way with some simple figures which are amenable to our purpose it may help to make clear the nature of the same problem in Statistics.

The following corresponding values of x and y were given in a Board of Education Examination (1911):—

$$x = 1{\cdot}00,\ 1{\cdot}50,\ 2{\cdot}00,\ 2{\cdot}30,\ 2{\cdot}50,\ 2{\cdot}70,\ 2{\cdot}80;$$
$$y = 0{\cdot}77,\ 1{\cdot}05,\ 1{\cdot}50,\ 1{\cdot}77,\ 2{\cdot}03,\ 2{\cdot}25,\ 2{\cdot}42.$$

Allowing for errors of observation, it was desired to test if there was a relation between y and x of the type

$$y = a + bx^2 \quad . \quad . \quad . \quad (1)$$

In the first place, the shape of the curve obtained by plotting y against x, as in fig. (6), would, to the initiated, probably suggest a parabola, the equation of which is of type (1). In order to test its suitability we proceed to plot y against x^2, or, putting $x^2 = \xi$, we plot y against ξ. If equation (1) holds, then, in that case

$$y = a + b\xi \quad . \quad . \quad . \quad (2)$$

should also hold, and this, in (ξ, y) co-ordinates, represents a straight line. The result of plotting y against ξ should therefore be a number of points approximately in a straight line—we say ‘approximately’ to allow for errors of observation in the original data.

Now from the given statistics corresponding values of ξ and y are, since $\xi = x^2$:—

$$\xi = 1{\cdot}00,\ 2{\cdot}25,\ 4{\cdot}00,\ 5{\cdot}29,\ 6{\cdot}25,\ 7{\cdot}29,\ 7{\cdot}84\ ;$$
$$y = 0{\cdot}77,\ 1{\cdot}05,\ 1{\cdot}50,\ 1{\cdot}77,\ 2{\cdot}03,\ 2{\cdot}25,\ 2{\cdot}42\ ;$$

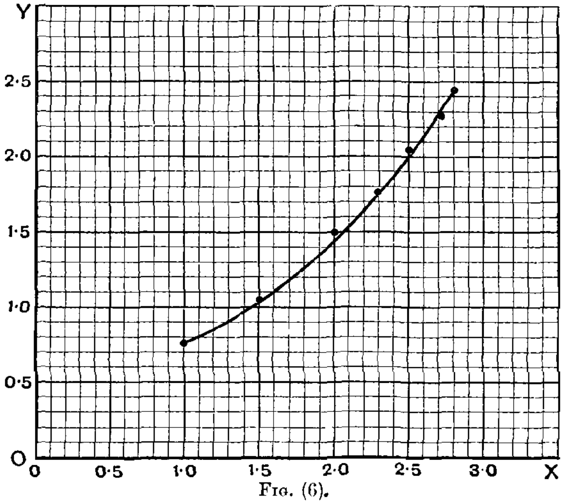


Fig. (6).

and the resulting graph, fig. (7), is very approximately a straight line. To determine its equation, choose two points (not too close together) on the line, which has been drawn so as to run as fairly as possible through the middle of the points plotted, and, in choosing, take points which lie at the *intersections* of horizontal and vertical cross lines (the printed lines of the graph paper) if such can be

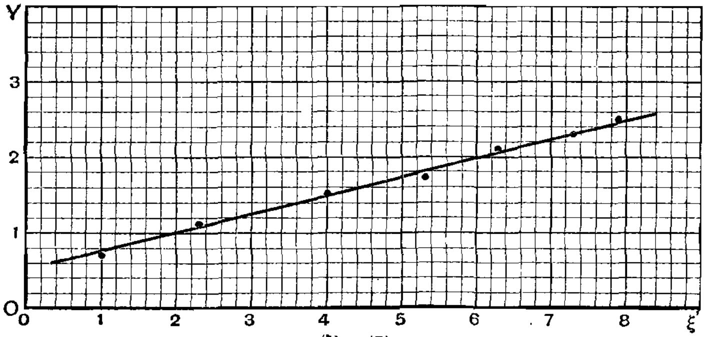


Fig. (7).

found, because their x's and y's can be read off with ease and accuracy. Two such points are

$$(2{\cdot}8,\ 1{\cdot}2) \text{ and } (6{\cdot}0,\ 2{\cdot}0),$$

and since each of these points lies on the line whose equation is

$$y=a+b\xi,$$

we have

$$1{\cdot}2=a+b(2{\cdot}8)$$
$$2{\cdot}0=a+b(6{\cdot}0).$$

Subtracting, we get

$$0{\cdot}8=b(3{\cdot}2).$$

Therefore $\qquad b=\frac{1}{4}.$

Hence $\qquad a=2-\frac{6}{4}=\frac{1}{2}.$

Thus the equation of the line is

$$y=\tfrac{1}{2}+\tfrac{1}{4}\xi,$$

i.e. $\qquad 4y=\xi+2,$

and the law connecting x and y is therefore

$$4y=x^2+2.$$

The following statistics, the result of an experiment in Physics to verify Boyle's Law, may be treated in the same way. x is a number proportional to the volume of a constant weight of gas in a closed space, and y is a number proportional to its absolute pressure. Corresponding values of x and y observed were :—

$x=$	46·89	41·96	40·33	38·88	37·37	36·06	34·71	33·47
$y=$	76·32	85·38	88·93	92·36	96·09	99·61	103·51	107·51

$x=$	32·39	31·08	29·97	28·76	27·26	25·32	24·04
$y=$	111·09	115·69	120·05	125·08	131·99	142·09	149·81.

Boyle's Law states that the product xy is constant, and this may be tested by putting $\xi=\frac{1}{x}$ and plotting y against ξ ; the points obtained should be approximately in a straight line.

Now in Statistics, as we have already explained, the exact connection between the variables, x and y, is rarely so clear, though the absence of law is not so complete as it might seem at first sight. At this stage, however, we need not enter into the difficult question of curve fitting : if drawn with care and used with judgment much that is of value may be learnt by simple plotting and by connecting up the resulting points by straight lines or a freehand curve. We shall briefly explain or illustrate by examples how graphs and

graphical ideas may be used to serve **three distinct purposes**, namely :—

(1) to suggest *correlation* or connection between two different factors or events ;
(2) to supply a basis for finding by *interpolation* some values of a variable when others are known ;
(3) as *pictorial arguments* appealing to the reason through the eye.

We reserve (2) and (3) for the next chapter and proceed at present with an example of (1).

Correlation suggested by Graphical means. Consider the index numbers, col. (2) Table (20), showing the variation from year to year in wholesale prices between the years 1871 and 1912. It is not an easy matter to take in satisfactorily the meaning of such a mass of bare figures, but they are much easier to grasp when plotted in a graph.

In this case the numbers x, representing years, and the numbers y, representing prices, are measures of things of quite a different character, so that it is not necessary to take the x and y units of the same size. Moreover they need not, in a case of this kind, necessarily vanish at the origin, but it is convenient to draw the graph in such a way that it shall occupy the greater part of the space at our disposal. Thus, we have roughly 80 small squares across the breadth of our graph paper, and between 1871 and 1912 we have roughly 40 years ; we therefore take two sides of a square to 1 year and mark off the years 1870, 1875, 1880, . . ., along an axis or base line parallel to the breadth of the paper, as shown in fig (8). Again we have roughly 70 small squares in the available space from this base line to the top of our graph paper, and the wholesale price index numbers vary from 88·2 to 151·9, a range of 63·7 ; we therefore take one side of a square to correspond to a difference of 1 in the price index number, and mark off the prices 90, 100, 110, . . . , along an axis parallel to the length of the paper, as shown in the figure.

We then plot points to represent the numbers in col. (2) of Table (20). Thus, in 1880 wholesale prices stood at 129 ; we therefore travel along the width of the paper till we reach 1880 and then upwards until we are *opposite* the 129 level on the axis of prices, inserting a dot to mark the position. Similarly for all other points, and the required graph is given by joining them up in succession.

Table (20). Marriage Rate and Wholesale Prices Index Numbers.

(1) Year.	(2) Prices.	(3) Nine Years' Average of Prices.	(4) Difference between Nos. in Cols. (2) & (3).	(5) Marriage rate.	(6) Nine Years' Average of Marriage rate.	(7) Difference between Nos. in Cols. (5) & (6).
1871	135·6	..	..	167	..	..
1872	145·2	..	..	174	..	..
1873	151·9	..	..	176	..	..
1874	146·9	..	..	170	..	..
1875	140·4	139·3	+1·1	167	164	+ 3
1876	137·1	138·6	−1·5	165	162	+ 3
1877	140·4	136·5	+3·9	157	159	− 2
1878	131·1	133·8	−2·7	152	157	− 5
1879	125·0	131·5	−6·5	144	155	−11
1880	129·0	128·5	+0·5	149	153	− 4
1881	126·6	125·2	+1·4	151	151	..
1882	127·7	120·8	+6·9	155	149	+ 6
1883	125·9	117·2	+8·7	155	148	+ 7
1884	114·1	114·7	−0·6	151	148	+ 3
1885	107·0	111·8	−4·8	145	149	− 4
1886	101·0	109·2	−8·2	142	149	− 7
1887	98·8	106·9	−8·1	144	149	− 5
1888	101·8	104·2	−2·4	144	149	− 5
1889	103·4	102·5	+0·9	150	149	+ 1
1890	103·3	101·0	+2·3	155	149	+ 6
1891	106·9	99·9	+7·0	156	150	+ 6
1892	101·1	98·7	+2·4	154	151	+ 3
1893	99·4	97·4	+2·0	147	153	− 6
1894	93·5	96·3	−2·8	150	155	− 5
1895	90·7	95·0	−4·3	150	156	− 6
1896	88·2	94·3	−6·1	157	156	+ 1
1897	90·1	93·8	−3·7	160	157	+ 3
1898	93·2	93·4	−0·2	162	158	+ 4
1899	92·2	93·8	−1·6	165	159	+ 6
1900	100·0	94·7	+5·3	160	159	+ 1
1901	96·7	95·7	+1·0	159	159	..
1902	96·4	96·9	−0·5	159	158	+ 1
1903	96·9	98·3	−1·4	157	158	− 1
1904	98·2	99·5	−1·3	153	156	− 3
1905	97·6	100·0	−2·4	153	155	− 2
1906	100·8	101·3	−0·5	157	154	+ 3
1907	106·0	102·8	+3·2	159	153	+ 6
1908	103·0	104·8	−1·8	151	153	− 2
1909	104·1	..	..	147	..	..
1910	108·8	..	..	150	..	..
1911	109·4	..	..	152	..	..
1912	114·9	..	..	155	..	..

It is comparatively easy from this graph to trace the change in prices from year to year and from decade to decade : for example, we note that from 1873 to 1896 the tendency of prices was on the whole downward, and from 1896 to 1910 the tendency was upward. Also on the assumption—not necessarily valid—that prices have varied continuously, or at least consistently, during the intervals between the dates to which the records refer, it is possible to read off intermediate values from the graph : *e.g.* midway between 1883 and 1884 we get the figure 120 as the index number for prices.

On the same graph sheet we have also plotted the marriage rate from year to year during the same period. The numbers are given in col. (5) of Table (20). This rate varies from 142 to 176, a range of 34, and we have a range of 40 small squares at our disposal in plotting ; a difference of 1 in the marriage rate has therefore been taken to correspond to one side of a square, and the marriage rates 140, 150, 160 . . . are accordingly marked along the axis perpendicular to the same base line as before, which is used again to measure the passage of years, but the second graph is drawn below the line whereas the first was drawn above it. In this way we are able to compare the two graphs, namely, the one registering the change in prices and the one registering the change in marriage rate from year to year.

It is interesting to observe that the two seem to be not unconnected : they go up and down almost in the same time, and mountains and valleys in the one correspond roughly to mountains and valleys in the other ; in other words, there is some kind of *correlation* or reciprocal relation between them. Now these mountains and valleys are largely the result of what may be called *short-time fluctuations*, and it is important to distinguish between these changes which are transient and the more permanent or *long-time changes*. In order to get rid of the former, which sometimes conceal the latter, the following device has been adopted : noticing that the wave period, the length of time taken for each complete up-and-down motion, is one of about nine years, nine-yearly averages have been taken of the figures for wholesale prices right down col. (2) of Table (20) ; thus 139·3 is the average of the index numbers from 1871 to 1879 inclusive, 138·6 is the average of the numbers from 1872 to 1880 inclusive, and so on, the results being recorded in col. (3). When the points corresponding to these numbers are plotted we get the broken line in fig. (8) passing through the body of the original graph of prices and indicating its general trend in the course of years as separated from the temporary fluctuations.

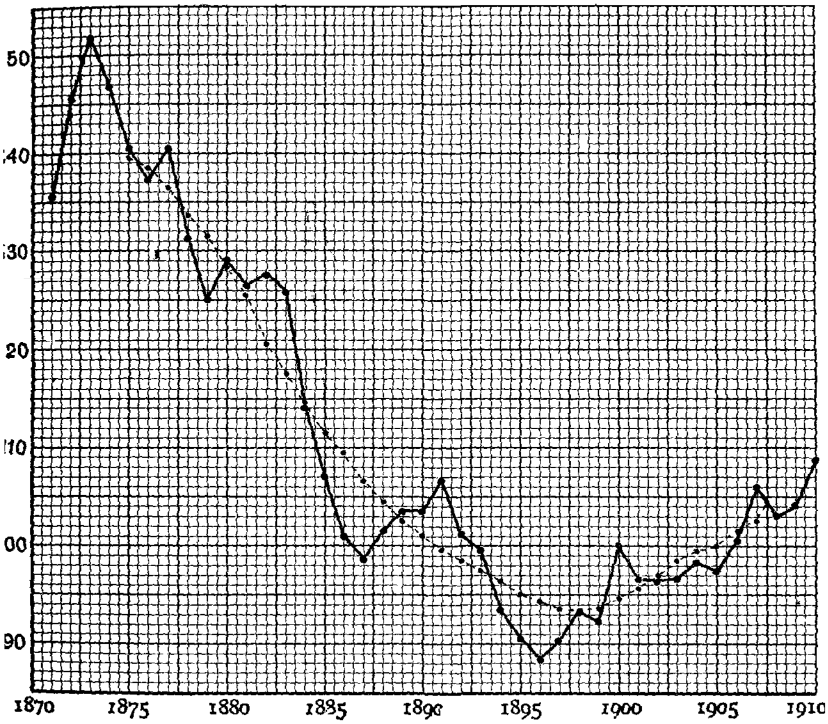


Fig. (8). Graph showing Variation in Wholesale Prices Index Numbers.

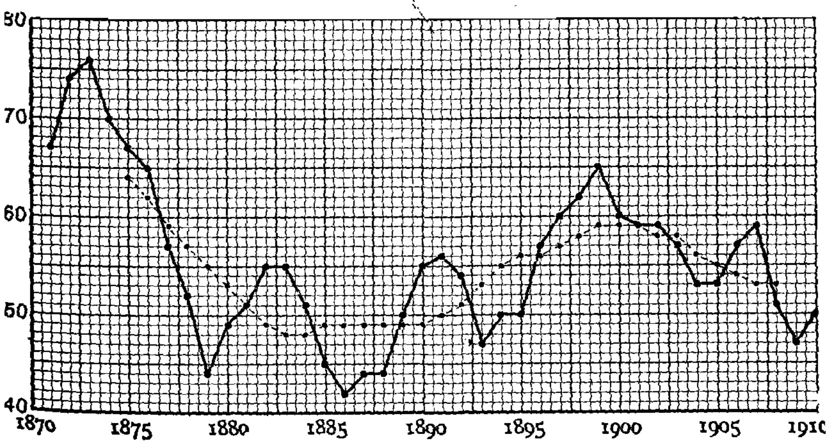


Fig. (9). Graph showing Variation in Marriage Rate Index Numbers.

The same procedure has been followed with the marriage rate statistics; the nine-yearly averages are shown in col. (6) of Table (20), and their graph appears as a broken line passing through the body of the original marriage rate graph in fig. (9).

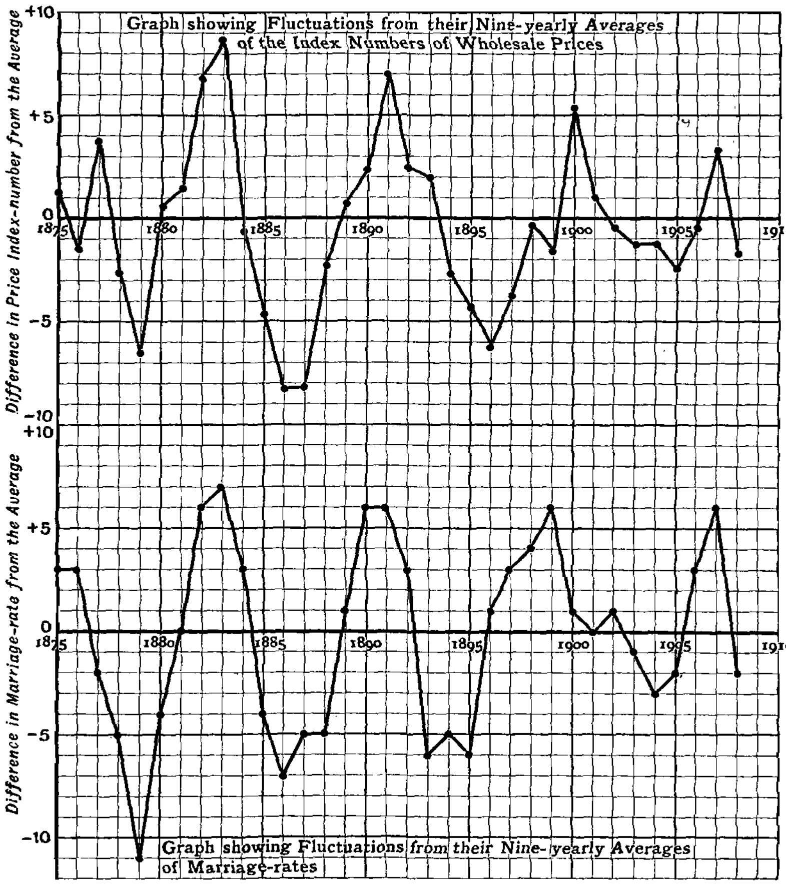


FIG. (10).

Suppose we wish on the other hand to study the short-time fluctuations as distinct from the 'secular trend,' we may do so by forming the differences between the numbers for each year and the corresponding nine-yearly averages, and plotting these differences on convenient scales.

The numbers obtained in this way are recorded, with their proper signs—positive if above the average, negative if below—in cols. (4) and (7) of Table (20), and the graphs of these differences are drawn,

one below the other for comparison, on the same graph sheet (fig. 10). The agreement in fluctuation from the average between the two factors, marriage rate and prices, is more easily remarked now than it was in the original graphs. High prices go as a rule hand-in-hand with prosperous times, and such times lead to more frequent marriages. This statement must not be taken to imply that when prices are high the times are always necessarily prosperous for the community as a whole : the lie direct would be given to such an implication by any one who had experienced abnormal war conditions.

After about 1892, while the fluctuations continue to be similar, a tendency appears for the marriage rate graph to reach each extreme point about a year in advance of the other, as though an increase in marriages raised prices and a decrease lowered them. There is no doubt that any economic change, especially if it takes place on a large scale, will set up a system of corresponding forces, sometimes in unexpected directions, actions and reactions succeeding one another at intervals like tidal waves producing each a backwash as it breaks, but such effects, even when anticipated in theory, are not always easy to unravel in practice.

The comparison we have been discussing between changes in prices and marriages is suggested in Sir W. H. Beveridge's *Unemployment.* The whole book will repay careful study, but it contains one particularly illuminating chapter on 'Cyclical Fluctuation' with a chart labelled 'The Pulse of the Nation,' because of the remarkable picture it gives of the ebb and flow of the tide of national prosperity. It consists of a series of curves representing respectively :—

(1) bank rate of discount per cent.;
(2) foreign trade as measured by imports and exports per head of the population ;
(3) percentage of trade union members not returned as unemployed ;
(4) number of marriages per 1000 of the population ;
(5) number of indoor paupers per 1000 of the population ;
(6) gallons of beer consumed per head of the population ;
(7) nominal capital of new companies registered in pounds per head of the population.

The interesting thing about these curves is to see the way in which they move in waves of varying size up and down almost together, showing a connection between such phenomena more

intimate than one might at first have suspected. A note of caution must be inserted here however : *causal* connection must not be too confidently inferred in discussing the correlation of characters changing simultaneously with time ; because two events happen together, one is not necessarily caused by the other.

An instructive article bearing on this point appeared recently in a periodical well known to students of social problems. It was there stated that high positive correlation exists between birth rate and infantile death rate : in general the two rise or fall together, whence Neo-Malthusians argue that the way to lower a death rate is to lower the birth rate. The writer then contrasts Bradford, the last word in the scientific care of infants, with Roscommon, where conditions as to wealth and child welfare are the very reverse, and points out that Bradford has a birth rate of 13 and an infant death rate of 135, while Roscommon has a birth rate of 45 and an infant death rate of 35. These figures, he suggests, prove instantaneously that the Neo-Malthusians are guilty of the commonest of all fallacies, they confound correlation with causation.

As an exercise in plotting the reader may see whether he can discover any suggestion of correlation between crime and unemployment by comparing the following statistics, showing the number of indictable offences tried in the United Kingdom and the trade union unemployed percentages respectively from 1861 to 1905 :—

TABLE (21). NUMBER OF TRIED INDICTABLE OFFENCES AND TRADE UNION UNEMPLOYED PERCENTAGES (1861-1905).

Year.	No. of Indictable Offences tried (in thousands).	Trade Union Unemployed percentages.	Year.	No. of Indictable Offences tried (in thousands).	Trade Union Unemployed percentages.
1861	56·0	3·7	1874	53·5	1·7
1862	61·3	6·0	1875	50·0	2·4
1863	61·4	4·7			
1864	58·4	1·9	1876	51·9	3·7
1865	59·9	1·8	1877	53·8	4·7
1866	57·6	2·6	1878	56·0	6·8
1867	59·5	6·3	1879	55·0	11·4
1868	62·4	6·7	1880	60·7	5·5
1869	61·3	5·9			
1870	56·1	3·7	1881	60·6	3·5
			1882	63·3	2·3
1871	53·1	1·6	1883	60·8	2·6
1872	51·9	0·9	1884	59·6	8·1
1873	53·5	1·2	1885	56·4	9·3

TABLE (21). NUMBER OF TRIED INDICTABLE OFFENCES AND TRADE UNION UNEMPLOYED PERCENTAGES (1861-1905)—*Continued.*

Year.	No. of Indictable Offences tried (in thousands).	Trade Union Unemployed percentages.	Year.	No. of Indictable Offences tried (in thousands).	Trade Union Unemployed percentages.
1886	56·2	10·2	1896	50·7	3·3
1887	56·2	7·6	1897	50·7	3·3
1888	58·5	4·9	1898	52·5	2·8
1889	57·6	2·1	1899	50·5	2·0
1890	55·0	2·1	1900	53·6	2·5
1891	54·1	3·5	1901	55·5	3·3
1892	58·3	6·3	1902	57·1	4·0
1893	57·4	7·5	1903	58·4	4·7
1894	56·3	6·9	1904	60·0	6·0
1895	50·8	5·8	1905	61·5	5·0

The chief point of difficulty in plotting such graphs is the initial one of fixing upon the most convenient scales to use, and in this matter hints only can be given, facility will come by practice. An examination of Table (21) shows that the data cover a period of forty-five years which can be marked off horizontally along a base line so as just to fit comfortably into the available space across the graph paper. The unemployed percentages vary between 0·9 and 11·4, giving a range of 10·5. Similarly the indictable offences recorded (in thousands) present a range of 13·3. We might therefore very well choose the same vertical scale for the measurement of indictable offences and unemployment, but, in order that the graphs may run more or less together (without exactly overlapping) for the sake of comparison, only the unemployment zero need be taken actually on the base line, whereas the indictable offences may have, say, the number 50 (thousand) at that level; also it will be convenient to show the scale for unemployment on the right side and the scale for offences on the left side of the paper.

An example dealing with matters somewhat different is provided by a comparison of changes from week to week in—

(1) the mean air temperature ;
(2) the percentage of possible sunshine ; and
(3) the rainfall.

The following is a record of observations taken at Greenwich in 1912 [data from *London Statistics,* vol. xxiii.] :—

TABLE (22). WEEKLY METEOROLOGICAL OBSERVATIONS AT GREENWICH (1912).

Week ended—	Mean Air Temperature—Degrees Fahrenheit.	Percentage of possible Sunshine.	Rainfall in inches.	Week ended—	Mean Air Temperature—Degrees Fahrenheit.	Percentage of possible Sunshine.	Rainfall in inches.
Jan. 6	45·7	7	0·76	July 6	58·7	15	0·36
13	41·9	15	0·45	13	67·0	46	0·20
20	40·2	1	0·93	20	65·8	44	0·04
27	38·9	8	0·88	27	64·8	31	0·16
Feb. 3	30·0	21	0·02	Aug. 3	57·8	33	0·54
10	39·5	15	0·52	10	57·6	28	1·26
17	45·5	11	0·44	17	56·2	14	0·23
24	47·4	6	0·65	24	57·2	24	1·27
Mar. 2	49·8	21	0·52	31	56·9	27	1·33
9	44·6	31	0·79	Sept. 7	54·8	36	0·21
16	45·1	16	0·19	14	52·4	14	0·02
23	42·7	15	1·08	21	53·6	22	0·00
30	51·0	46	0·05	28	51·5	59	0·02
Apr. 6	48·0	43	0·07	Oct. 5	48·8	36	2·30
13	45·6	43	0·02	12	46·0	53	0·00
20	50·0	50	0·00	19	49·8	38	0·13
27	52·6	76	0·00	26	45·4	23	0·88
May 4	50·1	32	0·21	Nov. 2	49·1	31	0·55
11	59·7	29	0·06	9	47·2	6	0·18
18	55·2	49	0·69	16	43·3	3	0·17
25	54·1	38	0·19	23	46·2	6	0·31
June 1	57·0	47	0·17	30	40·4	13	1·06
8	54·2	35	0·99	Dec. 7	42·4	9	0·31
15	58·1	48	0·39	14	49·0	2	0·62
22	61·7	56	0·65	21	44·4	19	0·59
29	60·2	45	0·30	28	48·1	8	1·22

The rainfall graph here should be drawn reversed (*i.e.* so that it goes up as the rainfall goes down in amount, and *vice versa*), because one would expect in general much rain to go with little sun and low temperature.

The range of temperature during the year is 37 degrees, of sunshine 75 per cent., and of rainfall 2·30 in. Hence the vertical scales for these three graphs might be chosen so that, roughly, 40 units of temperature should correspond to 80 units of sunshine and 2 units of rainfall. Also the zeros of the three variables should be so placed, relative to the horizontal base line registering the weeks, that the three graphs may be conveniently compared without causing confusion by too closely overlapping.

CHAPTER IX

GRAPHS (*continued*)

Graphical Ideas as a Basis for Interpolation. It frequently happens in statistical records that awkward gaps occur which require to be filled in; this may be due to the fact that no record has been made, or that it has been made with insufficient detail, or that it has been lost or destroyed. Cases in point arise in connection with returns like that of the Census which can only be undertaken every few years, so that if figures are wanted for any intervening year, as they are in very many instances, an estimate has to be made from the known results of the years recorded. It is imperative, for example, for many purposes of local or national government, to be able to find with a fair degree of accuracy the population of county boroughs and urban or rural districts at any given time, to know the number of workers engaged in different occupations, the amount of land in pasture and under various crops, the condition of the people as to housing, of the children as to education, and so on indefinitely.

Symbolically, with the same notation as we have used before, we conceive the statistics in tabular form, like

$$\left.\begin{matrix} x_1, x_2, x_3 & . & . & . & x_n \\ y_1, y_2, y_3 & . & . & . & y_n \end{matrix}\right\},$$

each y denoting the frequency corresponding to the character measured by its companion x, *e.g.* the x's may stand for successive dates and the y's for the frequencies of the population of a certain district at those dates. If it happens that one or more of the y's, in between the first and the last recorded, are missing, the problem is to estimate the missing values by some method of *interpolation*, as it is called. Various methods of arriving at such estimates are used, but we shall only refer to the more elementary here.

A rough way of making the estimate, but one which is often as accurate as the data will allow, is to plot the observations, each (x, y) being represented by a point, and connect them up, if there

be enough of them, by a smooth curve drawn freehand $P_1 P_2 P_3 \ldots P_n$ [see fig. (11)]; to find the y proper to any other x we have then only to draw the ordinate through the point (x, o) and measure the y at the point where it cuts the curve. This is a not unreasonable principle to follow, for in effect it gives due weight to each of the observations actually recorded, and it assumes an even course from each one to the next—a justifiable assumption in the absence of any evidence that some sudden discontinuity of value has taken place.

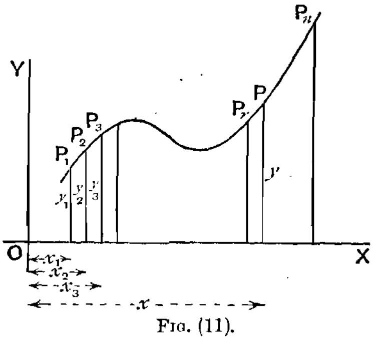


Fig. (11).

If only two observations are given, represented by the points P_1 (x_1, y_1) and P_2 (x_2, y_2), the curve connecting them is a straight line, and the y corresponding to any other x is at once given geometrically, as fig. (12) shows, by

$$\frac{PM}{P_2M_2}=\frac{P_1M}{P_1M_2}$$

i.e.

$$\frac{y-y_1}{y_2-y_1}=\frac{x-x_1}{x_2-x_1}$$

or

$$y=y_1+\frac{y_2-y_1}{x_2-x_1}(x-x_1),$$

the familiar proportional relation which is employed in this simple case.

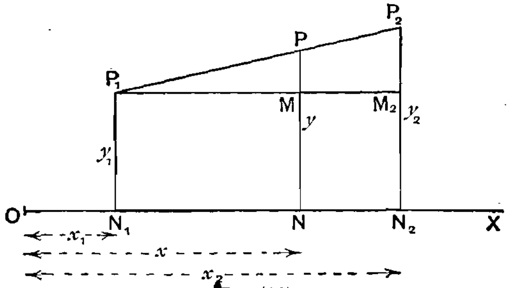


Fig. (12).

Example.—Given log 5·82673=0·7654249,
log 5·82674=0·7654257.

Required log 5·826736.

Here $x_1=5{\cdot}826730$ $y_1=0{\cdot}7654249$
$x_2=5{\cdot}826740$ $y_2=0{\cdot}7654257$
$x=5{\cdot}826736.$

Therefore, by means of the above relation,

$$y=0{\cdot}7654249+\frac{0{\cdot}0000008}{0{\cdot}000010}(0{\cdot}000006)$$
$$=0{\cdot}7654249+0{\cdot}00000048$$
$$=0{\cdot}7654254.$$

The logarithmic curve $y=\log x$ is, of course, not a straight line, and the value obtained for y only represents a first approximation to the true value.

When more than two points are given there is bound to be a margin of inaccuracy, more or less according to the data, introduced in drawing the curve. For an example of this method the *reader may refer back to the curve on p. 67, which was used to* determine the median and quartiles. We may, as we saw, read off from it the number of candidates who obtained not more than any stated number of marks: *e.g.* 300 candidates obtained not more than 34 marks; or we may use it the other way round and find the number of marks obtained by a stated number of candidates: *e.g.* 10 per cent. of the candidates got less than 17 marks. Such examples might be multiplied endlessly, and the method will be found extremely useful when a high degree of accuracy is not looked for. But greater confidence will be felt perhaps in such results—though the foundation for it may be no more secure in many cases—if we can translate them from geometrical to algebraical form, if we can find, that is to say, some formula, like the simple proportional relation already introduced above, which will give one y when others are known.

In order to make the argument as general as possible we shall speak of x and y as variables, and we shall think of the value of y as depending upon that of x in such a way that when x is given, y is known or it can be estimated * (in the sense that when the year is given the population is known or can be estimated).

Suppose

$$y=c_0+c_1x+c_2x^2+\quad .\quad .\quad .$$

[* This is equivalent to assuming that y is some function of x, say $y=f(x)$, and clearly some such assumption is necessary if any estimate from the known values to the unknown is to be possible. Further, for simplicity we assume $f(x)$ can be expanded in a Maclaurin's converging series of ascending powers of x, which simply means that we take the relation between x and y to be of the form adopted above.]

where the c's are constants to be determined, and their number can be made to depend upon the number of known values of y which are used in the estimate.

Geometrically, the equation

$$y=c_0+c_1x+c_2x^2+ \quad . \quad . \quad . \quad +c_nx^n$$

represents a curve called a parabola of the nth order, and such a curve could be employed (and uniquely found—there is only one parabola of the kind which will go through all the points) if we based our estimate upon a knowledge of $(n+1)$ y's corresponding to given x's, for we could readily make it pass through the $(n+1)$ known points (x_0, y_0), (x_1, y_1), (x_2, y_2), . . . (x_n, y_n) by choosing the $(n+1)$ c's so as to satisfy the $(n+1)$ simple linear relations :—

$$\begin{aligned} y_0&=c_0+c_1x_0+c_2x_0^2+ \quad . \quad . \quad . \quad +c_nx_0^n \\ y_1&=c_0+c_1x_1+c_2x_1^2+ \quad . \quad . \quad . \quad +c_nx_1^n \\ y_2&=c_0+c_1x_2+c_2x_2^2+ \quad . \quad . \quad . \quad +c_nx_2^n \\ &. \quad . \quad . \quad . \quad . \\ y_n&=c_0+c_1x_n+c_2x_n^2+ \quad . \quad . \quad . \quad +c_nx_n^n. \end{aligned}$$

When the curve is determined, in other words when the c's are known, we can find any other y required by substituting the corresponding x in the equation

$$y=c_0+c_1x+c_2x^2+ \quad . \quad . \quad . \quad +c_nx^n,$$

i.e. by supposing this point (x, y) to lie on the same curve that goes through the known points.

It is well to mention here that the parabola is by no means always the best curve for fitting any given statistics, and when the number of observations is adequate it is possible often to make a more satisfactory choice. Once the equation of a suitable curve has been determined the subsequent interpolation or calculation of y for any given x is not as a rule a very difficult matter. The larger question of curve fitting in general is reserved for a later chapter.

Example of First Method (fitting with a parabolic curve). Let us illustrate this process of interpolation by fitting a parabolic curve to the following figures, extracted from Porter's *The Progress of the Nation*, giving the annual cost of Poor Relief (excluding insane and casual) at five-yearly intervals, but with the amount for the year 1845 omitted :—

Year . . .	1835,	1840,	1845,	1850,	1855
Cost in £1000 . . .	5526,	4577,	?	5395,	5890

Assuming that no extraordinary conditions prevailed in 1845 to cause abnormality in expenditure, let us estimate what the figure would be for that year judging from the given records just before and after. Since there are four known points in this case, we take as the curve through them a parabola of the 3rd order, namely :—

$$y=c_0+c_1x+c_2x^2+c_3x^3\ ; \quad . \quad . \quad . \quad (1)$$

the four known points will then just suffice to determine uniquely the four arbitrary constants c_0, c_1, c_2, c_3. Also, since the x class-intervals are equal, it will simplify the algebra if we measure from the year 1845 as origin, taking five years as unit for x and £1000 as unit for y, so that we get

$$\left.\begin{array}{lllll} x=-2, & -1, & 0, & +1, & +2 \\ y=5526, & 4577, & y_0, & 5395, & 5890 \end{array}\right\}$$

where y_0 is the number to be determined.

Since all five points are to lie on the curve with equation as in (1), we have by substituting in that equation—

$$\begin{aligned} 5526&=c_0-2c_1+4c_2-8c_3 \\ 4577&=c_0-c_1+c_2-c_3 \\ y_0&=c_0 \\ 5395&=c_0+c_1+c_2+c_3 \\ 5890&=c_0+2c_1+4c_2+8c_3. \end{aligned}$$

Adding the first and last of these equations,

$$2c_0+8c_2=5526+5890 \quad . \quad . \quad . \quad . \quad (2)$$

Adding the second and last but one,

$$2c_0+2c_2=4577+5395$$

or $$8c_0+8c_2=4(4577+5395) \quad . \quad . \quad . \quad . \quad (3)$$

Subtracting (2) from (3),

$$\begin{aligned} 6c_0&=4(4577+5395)-(5526+5890) \quad . \quad . \quad (4) \\ &=4(9972)-(11416) \\ &=39888-11416 \\ &=28472. \end{aligned}$$

Therefore $\mathbf{y_0=c_0=£4,745,000.}$

If we only wish to make use of the records for the years 1840 and 1850, the appropriate fitting curve reduces to a straight line

$$y=c_0+c_1x,$$

on which we assume the points

$$(-1, 4577), \quad (0, y_0), \quad (+1, 5395)$$

to lie, so that

$$4577 = c_0 - c_1$$
$$y_0 = c_0$$
$$5395 = c_0 + c_1.$$

Therefore, adding the first and last of these equations,

$$2c_0 = 4577 + 5395,$$

so that $\mathbf{y_0 = c_0 = £4,986,000.}$

* *Second Method (using a formula connecting the ordinates).* When, as above, the steps from each x to the next are equal, as commonly happens in practice, it is possible to write down a simple relation between the y's, known and unknown, without introducing the c's at all. At bottom the method is the same as the last, inasmuch as the elimination of the c constants by the first method really results in the same formula for the unknown y.

Let us represent the given statistics in this case by

$$\left.\begin{array}{lllllll} x_0, & x_0+h, & x_0+2h & . & . & . & x_0+nh \\ y_0, & y_1, & y_2 & . & . & . & y_n \end{array}\right\}$$

so that, if the fitting curve be

$$y = c_0 + c_1 x + c_2 x^2 + \quad . \quad . \quad . \quad + c_n x^n,$$

we have, by substituting the co-ordinates of the first two points in this equation,

$$y_1 = c_0 + c_1(x_0+h) + c_2(x_0+h)^2 + \quad . \quad . \quad . \quad + c_n(x_0+h)^n$$

and $y_0 = c_0 + c_1 \, x_0 + c_2 \, x_0^2 + \quad . \quad . \quad . \quad + c_n x_0^n.$

Hence

$$y_1 - y_0 = c_1 h + c_2(2x_0 h + h^2) + \quad . \quad . \quad . \quad + c_n(n x_0^{n-1} h + \quad . \quad . \quad .).$$

Now this result, which we call *the* 1*st difference* between the y's, is of $(n-1)$th degree in x_0, so that by subtracting two of the y's we have reduced the degree in x_0 by 1. Similarly,

$$y_2 - y_1 = c_1 h + c_2(2x_0 h + 3h^2) + \quad . \quad . \quad . \quad . \quad + c_n(n x_0^{n-1} h + \quad . \quad . \quad .).$$

Thus we get a series of 1*st differences*, each with the highest term of the $(n-1)$*th degree in* x_0. Treating them as a series of new

[* The non-mathematical reader will do well to omit the rest of this section on

ordinates and forming their differences in the same way, we get what may be called *the 2nd differences* between the *y's*, a series of ordinates each with the highest term of *degree* $(n-2)$ *in* x_0. Proceeding in this way the *3rd differences* between the *y's* are a series of ordinates of *degree* $(n-3)$ *in* x_0, the *4th differences* are of *degree* $(n-4)$, and so on, until ultimately we reach the *nth differences,* which are of *zero degree in* x_0, and consequently involve only h. It follows that the nth differences must all be equal in value and therefore, if we go one step further and write down the $(n+1)th$ *differences,* these must *vanish altogether.*

If the reader finds any difficulty in following the argument he should test it step by step for himself in the simple case of a parabola of the third order when it should be perfectly clear.

The formation of the successive differences is conveniently shown in Table (23).

TABLE (23). SUCCESSIVE DIFFERENCES OF ORDINATES.

y	First difference Δ	Second difference Δ^2.	Third difference Δ^3.	Fourth difference Δ^4.	Fifth difference Δ^5.
y_0					
y_1	y_1-y_0				
y_2	y_2-y_1	$y_2-2y_1+y_0$			
y_3	y_3-y_2	$y_3-2y_2+y_1$	$y_3-3y_2+3y_1-y_0$		
y_4	y_4-y_3	$y_4-2y_3+y_2$	$y_4-3y_3+3y_2-y_1$	$y_4-4y_3+6y_2-4y_1+y_0$	
y_5	y_5-y_4	$y_5-2y_4+y_3$	$y_5-3y_4+3y_3-y_2$	$y_5-4y_4+6y_3-4y_2+y_1$	$y_5-5y_4+10y_3-10y_2+5y_1-y_0$

The law of formation should be apparent from this table, for it is precisely that which we meet in the binomial expansion, *e.g.* the nth difference is of type

$$y_n-ny_{n-1}+\frac{n(n-1)}{1\,.\,2}y_{n-2}-\frac{n(n-1)(n-2)}{1\,.\,2\,.\,3}y_{n-3}+\quad.\quad.\quad.\quad+(-1)^n y_0,$$

and by equating to zero the $(n+1)$th difference we have the relation required between the y's.

Example.—Let us apply this method to the ' Poor Relief ' example already considered. Since there are four known points the relation between x and y must be of the form

$$y=c_0+c_1x+c_2x^2+c_3x^3$$

as before. Hence the 4th differences must vanish, and taking the

points in order from years 1835 to 1855 as (x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), we get

$$y_4-4y_3+6y_2-4y_1+y_0=0$$

as the formula connecting five y's, four known and one (y_2) unknown.

Therefore
$$\begin{aligned}6y_2&=4(y_1+y_3)-(y_0+y_4)\\&=4(4577+5395)-(5526+5890),\end{aligned}$$

which is equivalent to equation (4) on p. 89.

Thus $\quad \mathbf{y_2=£4,745,000}$.

Third Method (by means of advancing differences). In the last method we employed a relation connecting y_n with all the preceding y's, but it is possible also to express y_n in terms of y_0 and the successive differences, which may be written $\triangle$, $\triangle^2$, $\triangle^3$, . . . $\triangle^n$; we have, in fact, with the notation of Table (23) :—

$$\triangle_0=y_1-y_0,\ \triangle_0^2=y_2-2y_1+y_0,\ \triangle_0^3=y_3-3y_2+3y_1-y_0,\ \ldots$$

Thus

$$\begin{aligned}y_1&=y_0+\triangle_0.\\y_2&=2y_1-y_0+\triangle_0^2=y_0+2\triangle_0+\triangle_0^2.\\y_3&=3y_2-3y_1+y_0+\triangle_0^3\\&=3(y_0+2\triangle_0+\triangle_0^2)-3(y_0+\triangle_0)+y_0+\triangle_0^3\\&=y_0+3\triangle_0+3\triangle_0^2+\triangle_0^3.\\y_4&=4y_3-6y_2+4y_1-y_0+\triangle_0^4\\&=4(y_0+3\triangle_0+3\triangle_0^2+\triangle_0^3)-6(y_0+2\triangle_0+\triangle_0^2)+4(y_0+\triangle_0)-\\&\quad y_0+\triangle_0^4\\&=y_0+4\triangle_0+6\triangle_0^2+4\triangle_0^3+\triangle_0^4.\end{aligned}$$

Here again the law of formation is clear, and it is readily established by induction that, *for all positive integral values of n,*

$$y_n=y_0+n\triangle_0+\frac{n(n-1)}{1.2}\triangle_0^2+\frac{n(n-1)(n-2)}{1.2.3}\triangle_0^3+\ \ldots, \quad . \quad . \quad (5)$$

a series which automatically comes to an end at the term $\triangle_0^n$.

An extension of this formula is obtained by writing θ in place of n, where $0<\theta<1$. We then get

$$y_\theta=y_0+\theta\triangle_0-\frac{\theta(1-\theta)}{1.2}\triangle_0^2+\frac{\theta(1-\theta)(2-\theta)}{1.2.3}\triangle_0^3-\ \ldots, \quad . \quad . \quad . \quad (6)$$

which enables us to interpolate for a y in between any two of a series of y's corresponding to x's advancing by equal steps. This relation

is no longer identically true as was (5), for the series on the right in (6) is unending, but its application in practice is justified when, as the differences advance, the numbers obtained tend to grow smaller and smaller, so that the remainder after a certain number of terms can be treated as negligible. Unless this tendency is realized without carrying the differences far the formula is not very satisfactory.

To illustrate the method of procedure the following figures may be used from Table (7), p. 25:—

TABLE (24). MARKS OBTAINED BY CERTAIN CANDIDATES IN AN EXAMINATION

No. of Marks.	No. of Candidates. y	First difference Δ	Second difference Δ^2	Third difference Δ^3
Not more than 45	447			
		37		
" " " 50	484		−16	
		21		1
" " " 55	505		−15	
		6		12
" " " 60	511		− 3	
		3		
" " " 65	514			

Suppose now we wish to know the number of candidates who obtained a number of marks not more than 48. In that case, in applying formula (6), we have

$$y_0=447, \qquad \theta=(48-45)/(50-45)=3/5,$$
$$\Delta_0=37, \quad \Delta_0{}^2=-16, \qquad \Delta_0{}^3=1,$$

and hence, up to this order of differences, the required number of candidates is given by

$$447+\tfrac{3}{5}\,.\,37-\frac{\frac{3}{5}(\frac{2}{5})}{1\,.\,2}(-16)+\frac{\frac{3}{5}(\frac{2}{5})(\frac{7}{5})}{1\,.\,2\,.\,3}(1)$$
$$=447+22{\cdot}2+1{\cdot}92+0{\cdot}06$$
$$=471, \text{ approximately.}$$

Also, number of candidates obtaining more than 48 marks, but not more than 50

$$=484-471$$
$$=13, \text{ approximately.}$$

Fourth Method (by means of Lagrange's Formula). We shall consider one more formula, due to the famous French mathematician Lagrange (1736-1813), which is useful when the recorded y's correspond to x's which advance by unequal stages.

Let the given statistics be represented as before by

$$(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots (x_n, y_n),$$

and consider the equation

$$y = y_0 \frac{(x-x_1)(x-x_2) \ldots (x-x_n)}{(x_0-x_1)(x_0-x_2) \ldots (x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2) \ldots (x-x_n)}{(x_1-x_0)(x_1-x_2) \ldots (x_1-x_n)} + \ldots + y_n \frac{(x-x_0)(x-x_1) \ldots (x-x_{n-1})}{(x_n-x_0)(x_n-x_1) \ldots (x_n-x_{n-1})} \quad \ldots (7)$$

It is of the nth degree in x, and it is identically satisfied by the $(n+1)$ pairs of values

$$(x=x_0, y=y_0), (x=x_1, y=y_1), \ldots (x=x_n, y=y_n).$$

It will therefore clearly serve as the fitting curve

$$y=c_0+c_1x+c_2x^2+ \ldots +c_nx^n,$$

being exactly of this type, and in order to get the y corresponding to any other x we have only to substitute that value of x in (7).

Example.—The following figures, based upon data from Porter's *The Progress of the Nation*, show the age distribution of criminals in the year 1842.

Percentage of criminals up to age 25=52·0 (y_0).
,, ,, ,, 30=67·3 (y_1).
,, ,, ,, 40=84·1 (y_2).
,, ,, ,, 50=92·4 (y_3).

Let us employ Lagrange's formula to find the approximate percentage of criminals up to 35 years of age, making use of the four ordinates given, and taking $x=35$. We have

$$y=52\frac{(35-30)(35-40)(35-50)}{(25-30)(25-40)(25-50)}+67\cdot3\frac{(35-25)(35-40)(35-50)}{(30-25)(30-40)(30-50)}$$
$$+84\cdot1\frac{(35-25)(35-30)(35-50)}{(40-25)(40-30)(40-50)}+92\cdot4\frac{(35-25)(35-30)(35-40)}{(50-25)(50-30)(50-40)}$$
$$=-10\cdot4+50\cdot475+42\cdot05-4\cdot62$$
$$=77\cdot5.$$

Reasoning made Clear with the Help of Graphs or Curves. The graphical method not only produces an instructive picture of a scheme of observations, but it may also be used effectively on occasion to pilot one through the intricacies of economic or similar argument. The eye is a very ready pupil and is quick to pass on what it sees to the mind; it acts, that is to say, as an ally to the understanding, which might get on without it, but which certainly gets on faster with it.

To illustrate this we shall consider the first principles of an interesting class of curves relating to supply and demand.*

Curve of Demand. Conceive a smoker who buys cigarettes at the rate of x per day, and pays for them at the rate of y pence each. Altogether they cost him therefore a sum of xy pence per day, which is conveniently measured by the rectangle OABC in fig. (13).

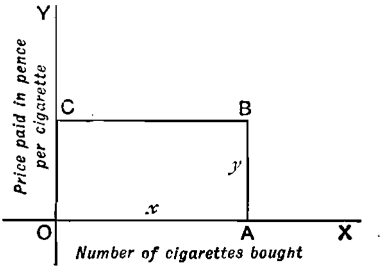


Fig. (13).

Notice that the cost price of each single cigarette is here represented by the area $(y \times 1)$, while the total expenditure is represented by the area $(y \times x)$.

Now let us suppose his country is at war and that the smoker, to put himself in a position to discourage luxuries, decides to give up smoking. Let us try to measure in terms of pence the cost of this great sacrifice to him on the first day.

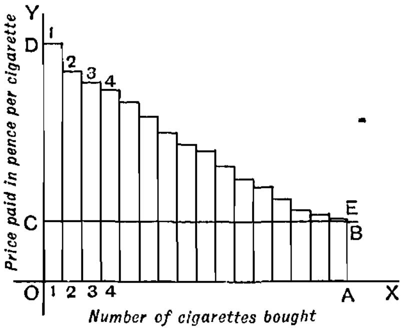


Fig. (14).

The first cigarette is probably the hardest to do without, and the desire for it is so strong that, if it were a mere matter of money and not of patriotism, he would be willing to give as many pence as are represented, say, by the rectangle 1–1 in fig. (14) in order to have it to smoke. If he went on to bargain

[* A fuller account of these curves will be found in Cunynghame's *Geometrical Political Economy*, where a rather more accurate interpretation of "surplus value" is given, involving the introduction of subordinate curves. The simplified statement here adopted seemed sufficient in an introductory course. Marshall's *Principles of Economics* also contains many fascinating illustrations of the use of such curves, mainly in footnotes.]

with himself in imagination, he would not be ready to offer quite so much for the satisfaction of a second smoke soon after the first: he would perhaps only give a number of pence represented by the rectangle 2–2 in the figure for this second cigarette. And if it came to a third he would offer less still, only '3–3' pence perhaps, for the fourth '4–4' pence, and so on. The rectangles here are of varying height, but each stands on a base of unit length.

Thus we find that the total sum he would be prepared to offer, bargaining for cigarette after cigarette in this way, would be represented by the sum of the rectangles 1–1, 2–2, 3–3 . . . in fig. (14), where the addition of each unit length along OX means one more cigarette in imagination smoked, and a diminution of unit length in an ordinate parallel to OY means a reduction of 1d. per cigarette in the price the smoker would be prepared to pay.

But if he fell a prey to his persistent craving and actually bought a number of cigarettes represented by OA in the figure, each would cost him in the ordinary way only a number of pence represented by AB, say, *i.e.* area (AB × 1), and his total expenditure would thus be measured by the area of the rectangle OABC. He would get them, that is to say, for less than he would be prepared to give rather than go without them. The difference, the area of the rectangles making up the portion BCDE of fig. (14), represents the measure in pence of surplus enjoyment which he would obtain free of charge, or it represents the measure of free sacrifice he makes if he is true to his patriotic principles.

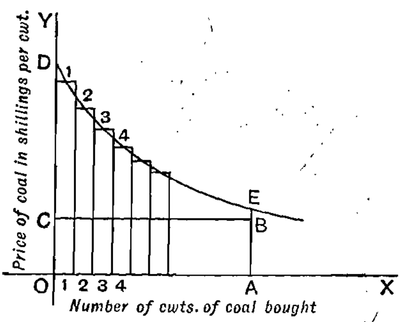


Fig. (15).

Let us now take an example on a larger scale. Imagine a small community of people, producers and consumers, buying and selling among themselves. Some of them are coalowners and sell coal to the others *in the open market, where competition is supposed free and unrestricted in any way.* This last condition is emphasized, because it is seldom perfectly satisfied in the real world of commerce.

Just as in the previous case we may represent the number of cwts. of coal bought by a length OA measured along OX in fig. (15), and the price actually paid in shillings per cwt. by the area of a rectangle on unit base and of height OC along OY. Thus the

total cost to the consumers in shillings is measured by the area of the rectangle OABC.

But here again we may picture the consumers during a coal shortage, when, rather than go without the first cwt. of coal, some one among them would be ready to offer for it as many shillings as are represented by the rectangle 1–1 in fig. (15), and for the second cwt. some one would be ready to offer '2–2' shillings, for the third '3–3' shillings, and so on. The demand for coal could thus be measured in shillings by the sum of the rectangles 1–1, 2–2, 3–3 . . . and, if OA runs into thousands of units of coal, the lengths 0–1, 1–2, 2–3 . . . along OX, corresponding to additions of 1 cwt. in the quantity bought, would in the limit be so small that the sum of the rectangles would become practically equivalent to the curvilinear area OAED in the figure, where DE is a curve drawn through the summits of the rectangles, namely the *curve of demand.*

The *consumers' surplus* in this case would be measured in shillings by the area BCDE, this being the difference between the measures of the sum actually paid for the coal bought and the sum consumers would have been willing to pay rather than go without it.

Curve of Supply. Now let us consider the question from the point of view of the coalowners. We shall assume that the average cost of production per cwt. of coal increases steadily as the number of cwts. produced increases; this would not be an unreasonable assumption in most cases after passing a certain point, since the richer coal measures known are likely to be mined before the poorer ones, and the cost of mining near the surface is bound to be less than when deep shafts have to be bored.

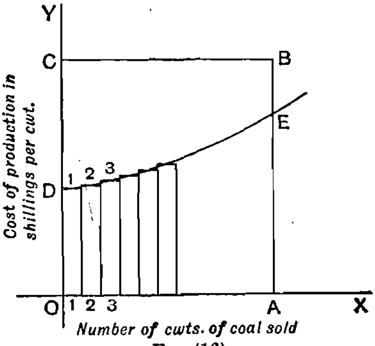


Fig. (16).

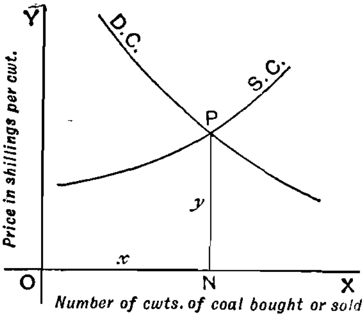
If, then, OA, fig. (16), represents the number of cwts. of coal sold, and if the price in shillings per cwt. at which it is sold is denoted by the area of a rectangle on unit base and of height OC along OY, the total payment received by the coalowners will be measured in shillings by the area of the rectangle OABC.

But the cost of producing the first cwt. is perhaps measured by the rectangle 1–1, that of producing the second cwt. by the rectangle 2–2, the third by the rectangle 3–3, and so on, each rectangle being drawn on unit base representing an advance of 1 cwt. (The

advance in the cost of production would not in reality be measured by so much the cwt. of course, but the assumption is inaccurate in degree only, not in principle, and, by making it, the argument is rendered clearer.) Thus the actual cost of production is, in the limit when OA is very large and divided up into relatively very small parts, measured in shillings by the curvilinear area OAED, where DE is a curve drawn through the summits of the rectangles namely, the *curve of supply*.

The difference, BCDE, between the areas OABC and OAED represents what is known as *producers' surplus*, for it measures the profit made by the owners in selling the coal at a higher price than the cost price of production.

Now let us combine the curve of supply (S.C.) and the curve of demand (D.C.) in the same figure, fig. (17). Their meeting point P determines the number of cwts. of coal bought (x), and the selling price in shillings per cwt. (y). For it is clear that under normal conditions it would not be profitable to coal producers to pass this point, because beyond it the demand on the part of coal consumers measured in money is less than the cost of production: they are not willing on the average to pay so much as ys. per cwt. for it, and it costs more than ys. per cwt. on the average to produce. If, on the other hand, the amount of coal produced decreases below x cwts., the greater this decrease the higher does the profit become on the sale of it, because the greater is the difference between the cost price and the selling price; hence, as profits become more pronounced, recruits will be attracted into the coal-producing business, and, if this goes on, deeper shafts will have to be bored and poorer fields worked until profits begin to decrease again and the supply once more approaches x cwts. Thus sooner or later the production of coal and its market price will tend to the level determined by the equilibrium point P where the supply and demand curves meet.



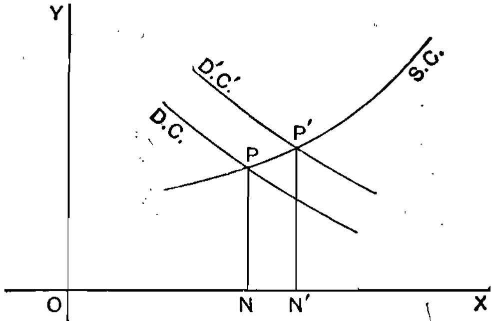
S.C. = *Supply curve*
D.C. = *Demand curve*

FIG. (17).

Endless varieties of problems may be discussed by altering the conditions and observing the effect produced in the standard diagram. Three examples will suffice to illustrate the method.

1. *Effect of a Change in Normal Demand.* Here we suppose the normal conditions of supply are unaltered—it costs just as much as before to produce the same amount of the commodity in question ; but a more eager demand on the part of consumers shows itself in a readiness to purchase more at any given price than would have been purchased under the old conditions : this may conceivably be due to a general increase in the purchasing power of these consumers, or it may be the result of a shortage of some other commodity which causes this one to be more widely used, just as margarine, for instance, has been known to take the place of butter ; whatever the reason may be, the effect is that the demand curve now occupies a higher level throughout its length, D′C′ in place of D.C. in the figures.

When we turn to the supply side of the question, there are three

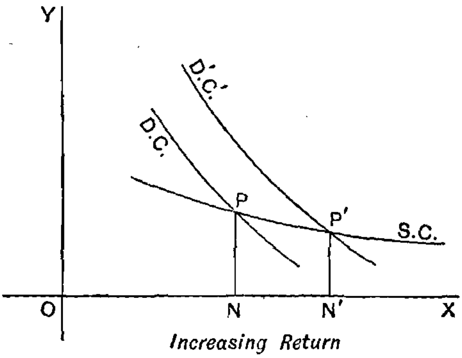


Decreasing Return

stages which, although they shade into one another in practice, it is well to separate clearly in theory : (1) the only supplies *immediately* available are those actually in the hands of dealers ; (2) to meet the increased demand, and so earn for themselves increased profits, manufacturers will speed up production, by working overtime, etc., with the help possibly of any disengaged labour or capital they may be able to secure, and the resulting extra supplies will be available *after a short time* ; (3) if the demand continues unabated manufacturers, by offering higher wages and interest, will seek to attract fresh labour and capital from other engagements into their business, and, by renewing their machinery and generally improving their organization, they will produce on a larger and relatively more economical scale. Moreover, other manufacturers, seeing the profits to be earned, will be attracted into the same line of business also, so that by this time the current available supplies of the commodity may exceed very appreciably their old figure.

But all this happens only *in the long run*, and the economist has always to bear this extremely important element of time carefully in mind when he seeks to estimate the effects of any proposed action.

We assume then that the new demand remains long enough at its higher level to allow for the gradual adjustment in this way of supply to the changed conditions, and for the economic forces called into play once again to arrive at a balance between them, most likely at a new equilibrium point. The two figures illustrate the difference in effect according as the production of the commodity is subject to a decreasing or an increasing return, *i.e.* according as the cost of production rises or falls when the amount produced is increased. In both cases it will be noted that more of the commodity is produced (ON′ in place of ON) in answer to the keener demand, but the difference is much greater in the second case than in the first. Also the price has gone up in the first case, while in the second it has gone down, the difference being measured by the change in PN.



2. *Effect of a Tax.* If the tax is at the rate of so much per unit (say 1s. per unit, if the price is measured in shillings) of the commodity produced, this will raise the supply curve, S.C., bodily up a distance of 1 unit into the position S′.C′., fig. (18), because the effect is the same as if 1s. were added to the cost of each unit in production. The production will thus be diminished by N′N units, for P′ is the new equilibrium point; the selling price will be increased by P′Ms per unit—by less, it should be noted, than P′Q or K′K, the amount of the tax; producers' surplus, which is analogous to what economists term

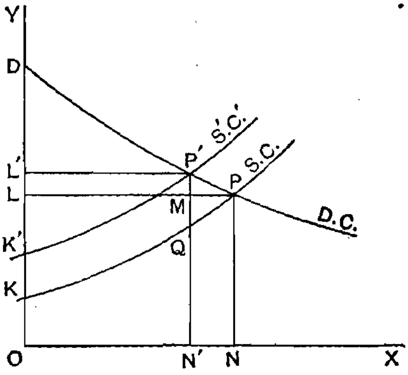


Fig. (18).

-ent, is diminished by (area KPL—area K′P′L′)s; consumers' surplus is diminished by (area PLL′P′)s; finally, the tax produces for the Treasury a number of shillings represented by a rectangle with sides of length ON′ and KK′.

3. *Effect of a Monopoly.* A monopolist has the power to stop production short of the true equilibrium point, so that ON′ cwts., fig. (19), are produced in place of the ON cwts. which free competition would demand. The selling price is thus raised by Q′Ss. per cwt.; producers' surplus is increased by (area KP′Q′M′—area KPL)s; while consumers' surplus is diminished by (area PLD—area DM′Q′)s.

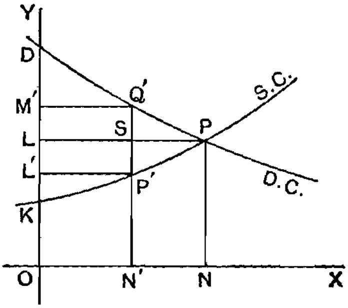


Fig. (19).

A word of explanation is necessary before leaving the subject of these supply and demand curves. It is probable that the reader will have questioned the possibility of drawing such curves for any commodity with sufficient accuracy to be of any value, but it would be enough as a rule to be able to estimate what would happen if a slight variation occurred in price or in production, and such an estimate may sometimes be made by actual trial: *e.g.* a good practical farmer most likely knows nothing about supply and demand curves as such, yet from past experience he has a pretty shrewd notion as to how far it may be profitable to spend an extra pound here in rearing calves and a pound less there in cultivating crops, bearing in mind the prices which cattle and corn might be expected to fetch. From his point of view the interest of the curves, if he knew anything of them, would be centred in those portions which correspond to normal conditions, *i.e.* somewhere in the neighbourhood of the equilibrium point under the free play of ordinary competition.

Their real value, however, as suggested at the beginning, does not consist in the practical assistance which they afford to the producer or consumer, by way of foretelling the actual measure of consumption or production, so much as in the light they throw upon general tendencies which are rather apt to be obscured if they are ponderously presented with elaborate economic argument. They make plain in a moment to the eye what can only be stated in two or three pages of writing.

CHAPTER X

CORRELATION

One of the most important questions which can be discussed by statistical methods is that of possible connection, or correlation, as it is called, between two sets of phenomena. If some factor in each can be isolated and measured numerically, our object is to discover if the size of either is sympathetically affected when a change occurs in the size of the other; or, to put the matter in another way, do large values of the one factor go with large values of the other factor and small with small, or *vice versa*? And, if some mutual dependence of this kind exists, can an estimate of its extent be made?

Consider, for example, the factor or character of height in husband and wife. Is there any connection between stature of husband (x) and stature of wife (y)? Do tall men tend on the average to wed tall women, or do we find tall men choosing short women for wives just about as often as they choose tall women? When correlation exists we shall want some measure for it which will tell us the amount of change or deviation from the average in either character associated with a given change or deviation from the average in the other.

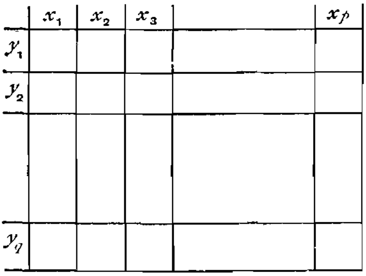


Fig. (20).

In studying graphs we saw how some hint of the existence of correlation might be discovered, but we wish now to go a little more deeply into the subject. The first step is to measure an adequate number of pairs of values, x and y, of the characters concerned in order to find what values are associated together, and how frequently the same values are repeated. When this is done we can draw up a table of double entry, see fig. (20), setting out in rows and columns the frequencies observed. An examination of Table (25), showing the variation of *brain weight* with *age*

in the case of 197 Bohemian women, will make clear what is meant. The x's from x_1 to x_p and the y's from y_1 to y_q are supposed to ascend in magnitude, and when, for example, the pair of values (x_2, y_3) is observed to be repeated nine times, the number 9 is placed in the second column and third row of the table, so that the frequency of each class is found recorded in the square proper to it: thus, out of the sample in Table (25), there are 10 women between the ages of 40 and 50 with brains weighing between 1300 and 1400 grams.

Table (25). Variation of Brain Weight with Age in the Case of certain Bohemian Women.

[Data from *Biometrika*, vol. iv. pp. 13 *et seq.*, *Variation and Correlation in Brain Weight*, by Raymond Pearl.]

		Age in years						
		x_1 20-30	x_2 30-40	x_3 40-50	50-60	60-70	x_p 70-80	*Totals*
Brain-weight in grams	y_1 1000-1100	1	-	1	1	-	-	3
	y_2 1100-1200	2	2	4	2	5	4	19
	y_3 1200-1300	28	9	8	14	10	4	73
	1300-1400	26	14	10	6	5	4	65
	1400-1500	13	7	7	2	-	2	31
	y_q 1500-1600	2	3	-	1	-	-	6
	Totals	72	35	30	26	20	14	197
	Mean y	1325	1350	1310	1285	1250	1279	

When each class interval, as in this table, includes a small range of values, the x and y may, as an approximation, be taken as the mid values of their class intervals: y_3 would be taken, for instance, as 1250, though it really includes all values between 1200 and

1300 grams. Strictly in such cases each single observation is not, geometrically speaking, located at a definite point, but lies somewhere within a small area, though it is treated as if it had the values x and y which apply to the centre point of the area. It is sometimes possible to correct for this assumption by what is known as Sheppard's adjustment, but we shall not concern ourselves with the correction in the present discussion, so as to avoid complications. because the difference made is not generally large.

The table, when drawn up, may immediately suggest some intimate connection between x and y. It may indicate that as x increases y also in general increases, or that y tends to fall in value as x grows bigger. But a more refined analysis is necessary. It would be instructive perhaps to travel along the row of x's, finding what mean value of y is associated with x_1, what mean value of y is associated with x_2, and so on. This would give a sounder basis for judging whether, as x increased, y in general increased or decreased as the case might be: for example, in Table (25) the mean values of y associated with the several types of x are shown in their proper columns at the foot of the table and clearly, as x increases, y tends to decrease, apart from conflicting readings at the beginning and end of the table, and the latter of these may not be significant of any real difference in brain weight at the end of life, for it is only based on fourteen observations; generally, the inference from this table would be that the weight of the brain decreases as the age increases after maturity is once reached, although, of course, it would be rash to make more than a tentative statement with so small a sample at our disposal.

Let us suppose $\bar{y}_1$ to be the mean value of y associated with x_1, $\bar{y}_2$ the mean value of y associated with x_2, $\bar{y}_3$ with x_3, and so on. If these values $(x_1, \bar{y}_1)$, $(x_2, \bar{y}_2)$, $(x_3, \bar{y}_3)$, etc., are plotted, it is very often found that they cluster more or less closely about a straight line, see fig. (21), so that we are led to ask whether there is not some line which will very fairly describe the run of the points; the equation of such a line would be

$$\bar{y}=mx+c,$$

and if m and c were known we could find from this equation the best average value of $\bar{y}$ corresponding to any given x.

But, on reflection, $\bar{y}_1, \bar{y}_2, \bar{y}_3 \ldots$ are themselves only the best y's corresponding to the particular values $x_1, x_2, x_3 \ldots$ of x, so that the problem is really the same as that of finding the relation

$$y=mx+c,$$

based on all the observations, which will enable us to estimate the best y corresponding to any given x.

Now for any value x_1 of x the value of y given by this relation is (mx_1+c), while by observation we may find more than one value of y corresponding to the value x_1 of x. If y_1 be one such value the difference between it and the value given by the above relation is

$$(mx_1+c)-y_1.$$

This difference we may regard as the error made in estimating y from the relation instead of taking the value given by observation

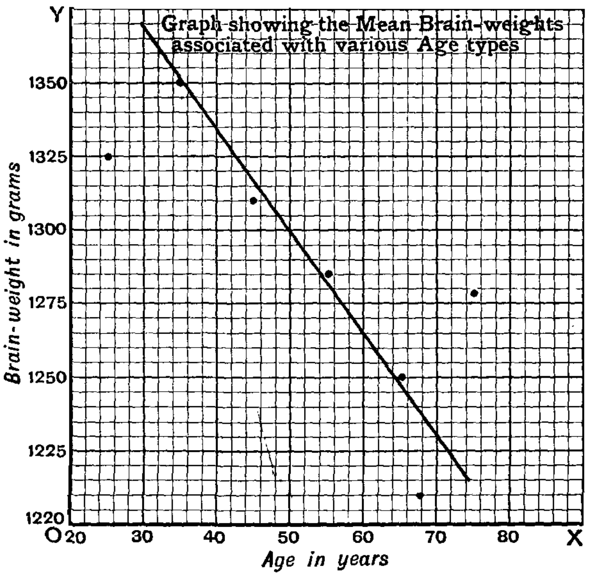


Fig. (21).

which for the moment we think of as the true value. The best relation will then clearly be the one which makes all such errors of estimate as small as possible. But, algebraically, some of these errors are positive, *i.e.* the value of y given by the relation is greater than that given by observation, and some are negative, and it is only their magnitudes that we wish to take into account. Accordingly we follow the method used in finding the standard deviation in order to get rid of the ambiguities of sign : we form, that is to say, the sum of the squares of the errors, because the expression so formed will clearly be least when each separate error is as small as possible in absolute magnitude.

To find, then, the values of m and c which will make

$$(mx_1+c-y_1)^2+(mx_2+c-y_2)^2+ \dots +(mx_n+c-y_n)^2$$

a minimum (see Note 7 in the Appendix), where n is the total number of pairs of observations.

The required values are given by differentiating, first with regard to c treating m as constant, and then with regard to m treating c as constant, putting each result equal to zero. Thus

$$(mx_1+c-y_1)+ \dots +(mx_n+c-y_n)=0$$

$$(mx_1+c-y_1)x_1+ \dots +(mx_n+c-y_n)x_n=0$$

Therefore $$m(x_1+ \dots +x_n)+nc-(y_1+ \dots +y_n)=0$$

$$m(x_1^2+ \dots +x_n^2)+c(x_1+ \dots +x_n)-(x_1y_1+ \dots x_ny_n)=0.$$

The first of these equations gives

$$m(n\bar{x})+nc-(n\bar{y})=0,$$

i.e. $$m\bar{x}+c-\bar{y}=0,$$

where $\bar{x}$ is the mean of all the x's and $\bar{y}$ is the mean of all the y's, and it expresses the fact that the line $y=mx+c$ passes through the point $(\bar{x}, \bar{y})$.

This might have been expected, for, graphically, each pair of observations (x_1, y_1), (x_2, y_2), (x_3, y_3) . . . corresponds to some point, and if we look for the line $y=mx+c$ passing through the region where they cluster most thickly together we should certainly expect it to pass through their mean or centre of gravity $(\bar{x}, \bar{y})$. This suggests how the values of m and c may be considerably simplified. If we measure all the x's from $\bar{x}$, their mean, and all the y's from $\bar{y}$, their mean, which is equivalent to taking the point $(\bar{x}, \bar{y})$ as origin and replacing every x by its deviation ξ from $\bar{x}$ and every y by its deviation η from $\bar{y}$, the first of the above relations is reduced to $c=0$, and therefore the second becomes

$$m(\xi_1^2+ \dots +\xi_n^2)-(\xi_1\eta_1+ \dots +\xi_n\eta_n)=0.$$

Hence $$m=(\xi_1\eta_1+ \dots +\xi_n\eta_n)/(\xi_1^2+ \dots +\xi_n^2)$$

$$=np/n\sigma_x^2$$

$$=p/\sigma_x^2,$$

where p is the mean of all the product pairs $\xi\eta$, and σ_x is the standard deviation of all the x's.

Thus the required equation for estimating the best η corresponding to any particular ξ is

$$\eta = p/\sigma_x^2 \cdot \xi,$$

whence

$$y - \bar{y} = \frac{p}{\sigma_x^2}(x - \bar{x}) \quad . \quad . \quad . \quad (1)$$

The coefficient p/σ_x^2 in this equation evidently gives the deviation in y from the mean y corresponding to unit deviation in x from the mean x, for when $(x - \bar{x}) = 1$, $(y - \bar{y}) = p/\sigma_x^2$. Hence the greater this coefficient is, the greater will be the change in y resulting from, or at all events coexistent with, unit change in x.

Thus p/σ_x^2 would seem to supply a not unreasonable measure of the correlation between x and y. But there is something very unsymmetrical about this result. Why should the correlation be measured by p/σ_x^2 any more than by p/σ_y^2? In fact, we might repeat the whole of the previous argument, interchanging x and y throughout wherever they appear. In that case we should first travel down the column of y's and calculate the mean values of x associated with y_1, y_1, y_3 . . . respectively. This would give a set of points $(\bar{x}_1, y_1)$, $(\bar{x}_2, y_2)$, $(\bar{x}_3, y_3)$, . . . , which, when plotted, would perhaps lie approximately in a straight line. We should thus be led to look for some relation

$$\bar{x} = m'y + c'$$

which would enable us to estimate the best average x corresponding to a y of given type, and, proceeding just as before, we should ultimately obtain the equation

$$\xi = p/\sigma_y^2 \cdot \eta,$$

or

$$(x - \bar{x}) = \frac{p}{\sigma_y^2}(y - \bar{y}), \quad . \quad . \quad . \quad (2)$$

in which the coefficient p/σ_y^2 gives now the deviation in x from the mean x corresponding to unit deviation in y from the mean y.

Hence p/σ_y^2 has, seemingly, just as much claim as p/σ_x^2 to measure the correlation between x and y. The one gives the change in x corresponding to unit change in y: the other gives the change in y corresponding to unit change in x; and the only reason why they differ is because unit change in x does not mean the same thing as unit change in y: their standards of changeableness or variability are not equal. If then we could alter the scales of measurement so that unit change in each were of the same magnitude, the two coefficients obtained ought to become identical, and we should then have a really satisfactory measure for the correlation required.

With this object let us examine the variability of the x's and compare it with the variability of the y's. Now the total dispersion of the different x's on either side of $\bar{x}$, the mean x, is conveniently measured by σ_x, their standard deviation. And similarly the dispersion of the y's on either side of $\bar{y}$, the mean y, is measured by σ_y. The bigger σ_x is, the greater is the variability of the x's, and the bigger σ_y is, the greater is the variability of the y's. Hence, in equations (1) and (2), $(x-\bar{x})$ should be divided by σ_x and $(y-\bar{y})$ by σ_y if we want to work with the same unit of change or variability in each case. The equations then become

$$\left(\frac{y-\bar{y}}{\sigma_y}\right)=\frac{p}{\sigma_x\sigma_y}\left(\frac{x-\bar{x}}{\sigma_x}\right)$$

and

$$\left(\frac{x-\bar{x}}{\sigma_x}\right)=\frac{p}{\sigma_x\sigma_y}\left(\frac{y-\bar{y}}{\sigma_y}\right).$$

Write $r=p/\sigma_x\sigma_y$; then r is taken to be the *coefficient of correlation*, for *it measures the change in either character corresponding to unit change in the other when the units are made comparable.*

The lines giving the best y for a given x and the best x for a given y may now be written

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

and

$$x-\bar{x}=r\frac{\sigma_x}{\sigma_y}(y-\bar{y}),$$

and they are called *lines of regression.* The term regression was first used by Sir Francis Galton in a paper entitled *Regression towards Mediocrity in Hereditary Stature,* though the root idea is not by any means confined to characters affected by heredity: it holds for any pair of correlated variables. Galton found that if a number of tall fathers are selected and their heights measured, the mean height being calculated, and if, further, the heights of the sons of these fathers are measured, their mean height being likewise calculated, the latter is not equal to the mean height of the selected fathers, but is rather nearer the mean height of the population as a whole. There is, that is to say, a regression or stepping back of the variable towards the general average. Professor Karl Pearson has remarked that 'in the existing state of our knowledge the recognition that the true method of approaching the problem of heredity is from the statistical side, and that the most we can hope at present to do is to give the *probable* character of the offspring

of a given ancestry, is one of the great services of Francis Galton to Biometry.'

The expressions $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are called *coefficients of regression*, and they register in the above particular case the amount of abnormality to be expected in the height of the sons when the amount of abnormality in the height of the fathers is known, and *vice versa*. The regression of the sons' height, y, on the fathers' height, x, is, in fact, defined as the ratio of the average deviation of the heights of the sons from the mean height of all sons to the deviation of the heights of the fathers from the mean height of all fathers, and hence it may be written

$$=(y-\bar{y})/(x-\bar{x})=r\sigma_y/\sigma_x.$$

To make the definition more general, instead of speaking merely in terms of height, we refer to any row or column—for there is no intrinsic difference between row and column—in a table like Table (25) as an *array* of y's or of x's, and selecting a particular *type*, say a particular value of x (like fathers of height x), we define the regression of the corresponding array of y's (like heights of sons of these fathers) on the type x to be the ratio of the average deviation of the array of y's from the mean y to the deviation of the selected type x from the mean x.

Example. To illustrate, let us take some figures due to Professor Pearson and Dr. Alice Lee [*Biometrika*, vol. ii. pp. 357 *et seq.*, *On the Laws of Inheritance in Man*]. Suppose the mean stature of all observed fathers, based on a sample of over 1000 observations $=67{\cdot}68$ in., with S.D.$=2{\cdot}70$ in.

Also suppose the mean stature of all sons$=68{\cdot}65$ in., with S.D. $=2{\cdot}71$ in., and that the correlation r between stature of father and stature of son$=0{\cdot}514$.

The regression of son on father as regards stature is then given by

$$(y-68{\cdot}65)=(0{\cdot}514)\frac{2{\cdot}71}{2{\cdot}70}(x-67{\cdot}68)$$

where x is the height of selected fathers and y the mean height of their sons.

Hence $$y=0{\cdot}516x+33{\cdot}73,$$

so that if we selected fathers of height 70 in., for example, the mean height of their sons would not be 70 in., but

$$(0{\cdot}516)(70)+33{\cdot}73=69{\cdot}85 \text{ in.},$$

i.e. there is a regression towards the general mean, 68·65 in., of all sons.

Also the coefficient of regression

$$=(0{\cdot}514)(2{\cdot}71)/(2{\cdot}70)$$
$$=0{\cdot}516.$$

It is not difficult to show that the greatest numerical value r can in general take is unity, for consider the expression for the sum of the squares of the differences between the observed deviations of the y characters from their mean and the corresponding deviations as deduced from the best fitting regression line,

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x}).$$

If, with our previous notation, η denote the *observed* deviation of the one character y, associated with a particular deviation, ξ, of the other character, x, then, since $(r\sigma_y/\sigma_x)\xi$ denotes the best value *given by the line,* the sum of the squares of the differences between these values

$$=\left(\eta_1-r\frac{\sigma_y}{\sigma_x}\xi_1\right)^2+\ .\ .\ .\ +\left(\eta_n-r\frac{\sigma_y}{\sigma_x}\xi_n\right)^2$$

$$=(\eta_1^2+\ .\ .\ .\ +\eta_n^2)-2r\frac{\sigma_y}{\sigma_x}(\xi_1\eta_1+\ .\ .\ .\ +\xi_n\eta_n)+r^2\frac{\sigma_y^2}{\sigma_x^2}(\xi_1^2+\ .\ .\ .\ +\xi_n^2)$$

$$=n\sigma_y^2-2r\frac{\sigma_y}{\sigma_x}(nr\sigma_x\sigma_y)+r^2\frac{\sigma_y^2}{\sigma_x^2}(n\sigma_x^2)$$

$$=n\sigma_y^2(1-r^2).$$

Since the sum of a number of squared quantities must be positive, it follows that r^2 must be less than 1 and hence r *lies between* -1 *and* $+1$.

Further, $n\sigma_y^2(1-r^2)$ can only vanish if every one of the squared quantities on the other side vanishes independently of the rest, so that we only get $r=\pm1$, when

$$\eta_1/\xi_1=\eta_2/\xi_2=\ .\ .\ .\ =\eta_n/\xi_n=r\sigma_y/\sigma_x.$$

In this case the deviation of the one character from its mean is always exactly proportional to the deviation of the other character from its mean, and the correlation is then said to be *perfect,* for it is equivalent to *causation.* In perfect correlation a one-to-one correspondence thus exists between the values of the two characters, for to one value of either there corresponds one and only one value of the other and the standard deviation of the array

(measuring its variability) corresponding to any selected type vanishes.

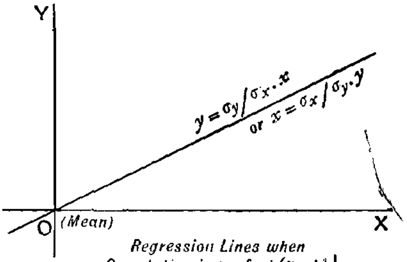
Zero correlation is at the opposite extreme where, no matter what the type selected in the one character may be, the mean value of the array in the second character is unaffected, because the two characters are quite independent or uncorrelated; the deviation of y from its mean bears no relation at all to the deviation of x from its mean, and unit change in either is associated with no particular change in the other, so that r must in this case be zero.

When r is negative, since $(y-\bar{y})/(x-\bar{x})=r\sigma_y/\sigma_x$ and the σ's are necessarily positive, corresponding to any value of x above the mean of all the x's the best value of $(y-\bar{y})$ is negative, that is, the best value of y is below the mean of all the y's, and *vice versa.* This means that in general high values of x would be associated with low values of y, and *vice versa.*

If we take the mean as origin so that the regression lines become

$$y=r\sigma_y/\sigma_x \,.\, x,$$
$$x=r\sigma_x/\sigma_y \,.\, y,$$

these lines coincide with the axes when the correlation is zero, and with one another when $r=\pm 1$ and the correlation is perfect, fig. (22). Given two equally variable characters ($\sigma_x=\sigma_y$) and perfect correlation, the regression lines coincide with one of the bisectors of the angle formed by the axes.



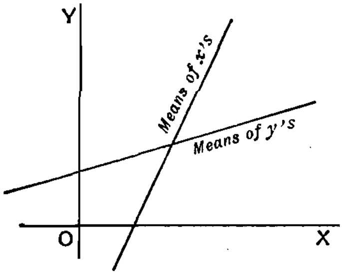
Regression Lines when Correlation is perfect (r=+1)

FIG. (22).

It may be helpful to look back again now at the graphical view of the argument leading up to the determination of the coefficient of correlation. For successive values of x we calculated the means of the several y's observed, these being presumably the best available y's corresponding to the particular x's selected, and we *assumed* that, when plotted, the points so obtained, $(x_1, \bar{y}_1)$, $(x_2, \bar{y}_2)$, $(x_3, \bar{y}_3)$, . . ., lay roughly in a straight line. In the same way we calculated the means of the several x's observed to correspond to particular y's selected, and again we *assumed* that the resulting points, $(\bar{x}_1, y_1)$, $(\bar{x}_2, y_2)$, $(\bar{x}_3, y_3)$ . . . lay roughly in a straight line. These assumptions are justified in very many cases, but when they fail recourse must be had to other methods beyond the scope of this book. [See

for example, Pearson's paper in *Drapers' Company Research Memoirs* Biometric Series II., *On the Theory of Skew Correlation and Non-linear Regression,* introducing the *correlation ratio,* η, which is equal to r in the particular case when the regression is linear.] Sometimes, again, although the observations are so scattered that the assumption of a straight line to describe the best fit seems somewhat wide of the mark, it may be justified on the ground that no better graphical result would be given by using any other curve in place of the line. Moreover the linear expression, $y=mx+c$, is simple and may serve to give at all events the first two terms of some more complex relation supplying an estimate for the most probable y corresponding to a given x.

If we had plotted *all* the original pairs of observations, instead of plotting certain x's and the *mean* y's associated with them, or certain y's and the associated *mean* x's, the two lines of regression would not have stood out so clearly: they would have lacked definition, like an optical image which is not strictly in focus, but there would have been a concentration of observations, as of light, in the neighbourhood where the lines of regression intersect, namely at $(\bar{x}, \bar{y})$, the mean of all the x's and all the y's. When, however, the lines of regression lie close together they become more clearly defined, all the observations being centred then more nearly in one line, and the correlation tends towards perfection. Such cases are frequent in Physics but rare, if found at all, in that class of Statistics into which the element of human impulse enters. When r is less than 1 the lines of regression, if the regression is of linear type, will be inclined to one another at some angle between 0 and 90 degrees.



If only a rough value of r, the correlation coefficient, is required, that may be obtained by merely estimating the gradient of each regression line and multiplying the results together, one measured relative to the axis of x and the other relative to the axis of y, for this product

$$=(\text{regression of } y \text{ on } x)\ (\text{regression of } x \text{ on } y)$$

$$=\left(r\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right)$$

$$=r^2.$$

Such an estimate may also be useful, though it may not be very dependable, when the complete distribution of characters is not known, for either regression line can be drawn when any two points on it are known and a single array of values of either character corresponding to a given type of the other is sufficient to fix one such point; also the mean $(\bar{x}, \bar{y})$, if it were known, would at once give a point common to both regression lines. When all the facts are available, however, the method of calculation is to be preferred to that of simply graphing the observations and their means, as there is bound to be a certain amount of guesswork and consequent error in deciding from a graph how the best regression lines run.

It is frequently convenient to refer the deviations of the given variables to some point other than the mean $(\bar{x},\bar{y})$ as origin, and, when this is done, a correction must be applied to the resulting value of r. We have already explained how, in such a case, to correct for standard deviations, and, as $r=p/\sigma_x\sigma_y$, it only remains to explain how to correct for p.

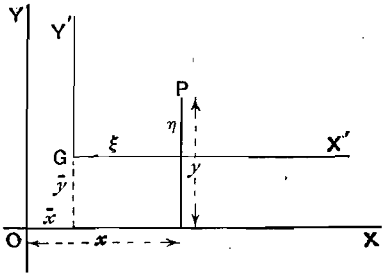


Fig. (23).

Now p is given by

$$np=\xi_1\eta_1+ \ . \ . \ . \ +\xi_n\eta_n,$$

where the ξ's and η's denote deviations from $\bar{x}$ and $\bar{y}$ respectively. Fig. (23) indicates the changes necessary in transferring from some origin O to the mean G. The co-ordinates of P (representing a typical observation) referred to O are (x, y) and referred to G are (ξ, η). Also the point G itself referred to O is $(\bar{x}, \bar{y})$. Thus

$$\xi=x-\bar{x}, \ \eta=y-\bar{y},$$

and np becomes

$$(x_1-\bar{x})(y_1-\bar{y})+ \ . \ . \ . \ +(x_n-\bar{x})(y_n-\bar{y})$$

$$=(x_1y_1-\bar{x}y_1-\bar{y}x_1+\bar{x}\bar{y})+ \ . \ . \ . \ +(x_ny_n-\bar{x}y_n-\bar{y}x_n+\bar{x}\bar{y})$$

$$=(x_1y_1+ \ . \ . \ . \ +x_ny_n)-\bar{x}(y_1+ \ . \ . \ . \ +y_n)-\bar{y}(x_1+ \ . \ . \ . \ +x_n)+n\bar{x}\bar{y}$$

$$=(x_1y_1+ \ . \ . \ . \ +x_ny_n)-\bar{x} \ . \ n\bar{y}-\bar{y} \ . \ n\bar{x}+n\bar{x}\bar{y}$$

$$=\Sigma(xy)-n\bar{x}\bar{y},$$

where $\Sigma(xy)$ denotes the sum of expressions of the type xy.

Hence the corrected value of p

$$=\Sigma(xy)/n-\bar{x}\bar{y},$$

from which we infer that the corrected value of r is

$$= \frac{\Sigma(xy) - n\bar{x}\bar{y}}{\sqrt{(\Sigma x^2 - n\bar{x}^2)(\Sigma y^2 - n\bar{y}^2)}}.$$

We proceed to a few applications of these results in the next chapter.

[As early as 1846 a French physicist, Auguste Bravais, had conceived the surface of error as a means of describing in space the path of a point whose x and y co-ordinates are subject to errors which are not independent; but it appears to be doubtful whether he saw the connection between his work and the subject of correlation. It was Galton, nearly forty years later, who really created that subject, introducing the coefficient of correlation on graphical lines and giving practical examples of its use. (See *Biometrika*, vol. xiii., pp. 25-45, *Notes on the History of Correlation.*)

Edgeworth, in 1892, using Galton's function, independently reached some of Bravais' results related to the correlation of three variables, and showed how they could be extended. Karl Pearson, in 1896, contributed to the Royal Society *Transactions* a fundamental paper on the subject, with special reference to the problem of heredity, drawing attention to the best value of the correlation coefficient, and how it should be calculated. (See Appendix, Note 11.) Yule, returning in the following year to Bravais' formulæ, showed their significance also in the case of skew correlation.

Pearson afterwards developed a method of determining the correlation of characters not quantitatively measurable, and in a discussion of the general theory of skew correlation in another paper he proposed a new function, the correlation ratio, applicable to the case of non-linear regression.]

CHAPTER XI

CORRELATION—EXAMPLES

Example (1).—To find the correlation between *Differences in Wholesale Price Index Numbers* and in the *Marriage Rate from their* corresponding *Nine-yearly Averages* during the twenty years, 1889-1908. using the data given on p. 77.

TABLE (26). CORRELATION BETWEEN DIFFERENCES IN WHOLESALE PRICES AND MARRIAGE RATE FROM THEIR RESPECTIVE NINE-YEARLY AVERAGES.

(1) Year.	(2) Difference in Prices from 9-yearly Average.	(3) Square of No. in Col. (2).	(4) Difference in Marriage-rate from 9-yearly Average.	(5) Square of No. in Col. (4).	(6) Product of Nos. in Col. (2) and Col. (4).
	(x)	(x^2)	(y)	(y^2)	(xy)
1889	+ 0·9	0·81	+ 1	1	+ 0·9
1890	+ 2·3	5·29	+ 6	36	+ 13·8
1891	+ 7·0	49·00	+ 6	36	+ 42·0
1892	+ 2·4	5·76	+ 3	9	+ 7·2
1893	+ 2·0	4·00	− 6	36	−12·0
1894	− 2·8	7·84	− 5	25	+ 14·0
1895	− 4·3	18·49	− 6	36	+ 25·8
1896	− 6·1	37·21	+ 1	1	− 6·1
1897	− 3·7	13·69	+ 3	9	−11·1
1898	− 0·2	0·04	+ 4	16	− 0·8
1899	− 1·6	2·56	+ 6	36	− 9·6
1900	+ 5·3	28·09	+ 1	1	+ 5·3
1901	+ 1·0	1·00	..	..	..
1902	− 0·5	0·25	+ 1	1	− 0·5
1903	− 1·4	1·96	− 1	1	+ 1·4
1904	− 1·3	1·69	− 3	9	+ 3·9
1905	− 2·4	5·76	− 2	4	+ 4·8
1906	− 0·5	0·25	+ 3	9	− 1·5
1907	+ 3·2	10·24	+ 6	36	+ 19·2
1908	− 1·8	3·24	− 2	4	+ 3·6
	+24·1 −26·6	197·17	+41 −25	306	+141·9 −41·6

The arithmetic is comparatively simple in this case because there is only one value of each variable corresponding to each year, so that there is no weighting or grouping to complicate the analysis. The variables x and y, between which we wish to find the correlation, appear in col. (2) and col. (4) in Table (26), and the positive and negative differences are separated from one another in each case so as to make their summation easier.

Thus for the arithmetic mean of the numbers in col. (2), we have

$$\bar{x}=(+24\cdot1-26\cdot6)/20=-0\cdot125\ ;$$

and for the mean of the numbers in col. (4), we have

$$\bar{y}=(+41-25)/20=+0\cdot8.$$

The straightforward procedure would now be to get the twenty corresponding values of ξ and η, the deviations of the twenty x's in col. (2) and of the twenty y's in col. (4) from $\bar{x}$ and $\bar{y}$ respectively, and, having found σ_x and σ_y, we could immediately deduce r from the formula

$$\begin{aligned} r&=p/\sigma_x\sigma_y \\ &=(\xi_1\eta_1+\ .\ .\ .\ +\xi_{20}\eta_{20})/20\sigma_x\sigma_y. \end{aligned}$$

But it is simpler to measure the deviations from (0, 0) as origin rather than from the mean ($-0\cdot125$, $+0\cdot8$), because x^2, y^2, and xy involve fewer significant figures than would ξ^2, η^2, and $\xi\eta$, and, of course, it will be necessary to correct for this at the end in the usual way.

The mean square deviation of x referred to zero as origin

$$=197\cdot17/20, \text{ by col. (3).}$$

Therefore, $$\begin{aligned} \sigma_x^2&=197\cdot17/20-(0\cdot125)^2=9\cdot843 \\ \sigma_x&=\mathbf{3\cdot14}. \end{aligned}$$

Again, the mean square deviation of y referred to zero as origin

$$=306/20, \text{ by col. (5).}$$

Therefore, $$\begin{aligned} \sigma_y^2&=306/20-(0\cdot8)^2=14\cdot66 \\ \sigma_y&=\mathbf{3\cdot83}. \end{aligned}$$

Also the corrected p

$$\begin{aligned} &=(\Sigma xy)/n-\bar{x}\bar{y} \\ &=100\cdot3/20-(-0\cdot125)(+0\cdot8), \text{ by col. (6)} \\ &=5\cdot015+0\cdot100 \\ &=\mathbf{5\cdot115}. \end{aligned}$$

Hence $$\begin{aligned} r&=p/\sigma_x\sigma_y \\ &=5\cdot115/(3\cdot14)(3\cdot83) \\ &=\mathbf{0\cdot43}. \end{aligned}$$

It is necessary to be careful with the signs in forming the numbers in col. (6), but otherwise the actual calculation should present no difficulty.

The regression equation giving the best marriage rate difference, Y, for a given wholesale price difference, X, from their respective nine-yearly averages is

$$(Y-0{\cdot}8)=r\frac{\sigma_y}{\sigma_x}\,.\,(X+0{\cdot}125)$$

$$=(0{\cdot}43)\frac{(3{\cdot}83)}{(3{\cdot}14)}(X+0{\cdot}125)$$

i.e. $\mathbf{Y=0{\cdot}52X+0{\cdot}86.}$

The regression equation giving the best wholesale price difference, X, for a given marriage rate difference, Y, from their respective nine-yearly averages is

$$(X+0{\cdot}125)=r\frac{\sigma_x}{\sigma_y}\,.\,(Y-0{\cdot}8)$$

$$=0{\cdot}35(Y-0{\cdot}8)$$

i.e. $\mathbf{X=0{\cdot}35Y-0{\cdot}40.}$

We noted that fig. (10), p. 80, suggested a closer correlation between the two factors we have been considering during the earlier years of the period 1875-1908 than during the later years. It might be worth while as an exercise to see if this is borne out by calculating r for the years 1875-1889, and comparing it with the value found for the years 1889-1908.

Example (2).—To find the correlation between *Overcrowding* and *Infant Mortality* in *London Districts.* [Data taken from *London Statistics,* vol. 23, published by the London County Council.]

The figures are apparently based upon the Census Report of 1911. The numbers in col. (2), Table (27), show what percentage of the total population occupying private houses in each district were living in overcrowded conditions, any ordinary tenement which has more than two occupants to a room, including bedrooms and sitting-rooms, being defined as overcrowded. The numbers in col. (5) show the infantile mortality in each district, that is, the number of infants who died under one year out of every 1000 born, including both sexes.

For the sake of comparison these numbers have been plotted together on the same graph sheet. The districts, arranged in alphabetical order, were numbered from 1 to 29 so as to form a horizontal scale corresponding to the scale of years in discussing prices and marriages. The scale in this case is, of course, purely artificial,

and the only reason for joining up neighbouring points is that we are better able by so doing to see whether or not high values of the one variable go with high values of the other variable, and low with low.

In calculating the mean and standard deviation for overcrowding we have measured deviations from 17·0 as origin, and in making the same calculations for infant mortality we have measured deviations from 125 as origin. It is convenient, therefore, to use the point (17·0, 125) as origin in working out also the product deviation sum, col. (8) of Table (27), instead of using the mean (17·86, 126).

TABLE (27). CORRELATION BETWEEN OVERCROWDING AND INFANT MORTALITY IN LONDON DISTRICTS (1911).

(1) District.	(2) Percentage of Population Overcrowded	(3) Deviation of No. in Col. (2) from 17·0.	(4) Square of No. in Col. (3).	(5) Infant Mortality.	(6) Deviation of No. in Col. (5) from 125.	(7) Square of No. in Col. (6).	(8) Product of Nos. in Col. (3) and Col. (6).
		(x)			(y)		
) Battersea . .	13·3	− 3·7	13·69	124	− 1	1	+ 3·7
) Bermondsey .	23·4	+ 6·4	40·96	156	+ 31	961	+ 198·4
) Bethnal Green .	33·2	+ 16·2	262·44	151	+ 26	676	+ 421·2
) Camberwell .	13·5	− 3·5	12·25	109	− 16	256	+ 56·0
) Chelsea . .	14·9	− 2·1	4·41	109	− 16	256	+ 33·6
) City of London	12·3	− 4·7	22·09	124	− 1	1	+ 4·7
) Deptford . .	12·2	− 4·8	23·04	142	+ 17	289	− 81·6
) Finsbury . .	39·8	+ 22·8	519·84	156	+ 31	961	+ 706·8
) Fulham . .	14·6	− 2·4	5·76	125	...	...	...
) Greenwich .	12·1	− 4·9	24·01	128	+ 3	9	− 14·7
) Hackney . .	12·4	− 4·6	21·16	119	− 6	36	+ 27·6
) Hammersmith .	14·2	− 2·8	7·84	146	+ 21	441	− 58·8
) Hampstead .	7·1	− 9·9	98·01	78	− 47	2209	+ 465·3
) Holborn . .	25·6	+ 8·6	73·96	115	− 10	100	− 86·0
) Islington . .	20·0	+ 3·0	9·00	127	+ 2	4	+ 6·0
) Kensington .	17·1	+ 0·1	0·01	133	+ 8	64	+ 0·8
) Lambeth . .	13·6	− 3·4	11·56	123	− 2	4	+ 6·8
) Lewisham. .	3·9	− 13·1	171·61	104	− 21	441	+ 275·1
) Paddington .	16·2	− 0·8	0·64	127	+ 2	4	− 1·6
) Poplar . .	20·6	+ 3·6	12·96	157	+ 32	1024	+ 115·2
) St. Marylebone	20·7	+ 3·7	13·69	108	− 17	289	− 62·9
) St. Pancras .	25·5	+ 8·5	72·25	112	− 13	169	− 110·5
) Shoreditch .	36·6	+ 19·6	384·16	170	+ 45	2025	+ 882·0
) Southwark .	25·8	+ 8·8	77·44	144	+ 19	361	+ 167·2
) Stepney .	35·0	+ 18·0	324·00	144	+ 19	361	+ 342·0
) Stoke Newington	8·8	− 8·2	67·24	102	− 23	529	+ 188·6
) Wandsworth .	6·3	− 10·7	114·49	122	− 3	9	+ 32·1
) Westminster .	12·9	− 4·1	16·81	103	− 22	484	+ 90·2
) Woolwich. .	6·3	− 10·7	114·49	97	− 28	784	+ 299·6
		+ 119·3 − 94·4	2519·81		+ 256 − 226	12748	+ 4322·9 − 416·1

For overcrowding,

$$\text{mean}=17+24\cdot9/29=17\cdot86\ ;$$

$$\sigma_x=\sqrt{[(2519\cdot81/29)-(0\cdot86)^2]}=\sqrt{(86\cdot15)}=9\cdot3.$$

For infant mortality,

$$\text{mean}=125+30/29=126\cdot03\ ;$$

$$\sigma_y=\sqrt{[(12748/29)-(1\cdot03)^2]}=\sqrt{438\cdot5}=20\cdot9.$$

Also p, referred to (17·0, 125)=(4322·9−416·1)/29=3907/29, and, referred to the mean (17·86, 126·03), this becomes

$$=3907/29-(0\cdot86)(1\cdot03)$$
$$=133\cdot8.$$

Hence $$r=133\cdot8/(9\cdot3)(20\cdot9)=\mathbf{0\cdot69},$$

so that the correlation between overcrowding and infant mortality is fairly marked.

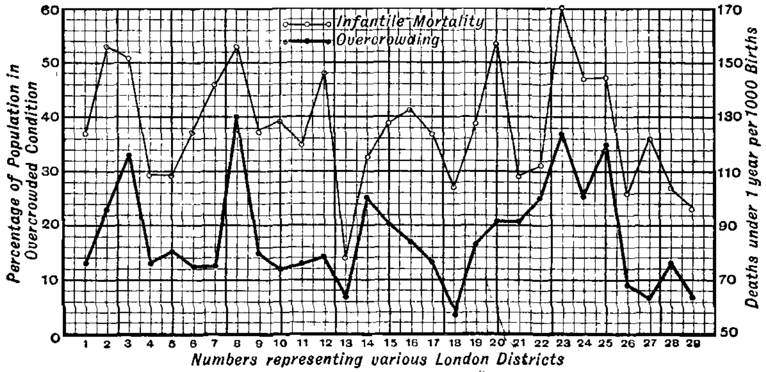


FIG. (24).

The regression equation giving the average infant mortality, Y, for districts in which the extent of overcrowding, X, is known is

$$Y-126\cdot03=r\frac{\sigma_y}{\sigma_x}(X-17\cdot86)$$

$$=\frac{(0\cdot69)(20\cdot9)}{9\cdot3}(X-17\cdot86)$$

i.e. $$\mathbf{Y=1\cdot55X+98\cdot4.}$$

Similarly, the regression equation giving the average percentage of overcrowding, X, for districts with a known amount of infant mortality, Y, is

$$X-17\cdot86=r\frac{\sigma_x}{\sigma_y}(Y-126\cdot03)$$

$$=0\cdot31(Y-126\cdot03)$$

i.e. $$\mathbf{X=0\cdot31Y-21\cdot0.}$$

Example (3).—The reader might apply the same method to the determination of the correlation between *Ratio of Indoor Paupers* and *Ratio of Outdoor Paupers*, each measured *per* 1000 *of the* estimated *Population* in *England and Wales*, excluding casuals and insane, during the years 1900-1914. The following are the statistics required for the purpose :—

TABLE (28). CORRELATION BETWEEN RATIO OF INDOOR AND RATIO OF OUTDOOR PAUPERS, EACH MEASURED PER 1000 OF THE POPULATION.

Year.	Indoor Paupers—Rate per 1000.	Outdoor Paupers—Rate per 1000.	Year.	Indoor Paupers—Rate per 1000.	Outdoor Paupers—Rate per 1000.
1900	5·9	15·8	1908	6·8	15·4
1901	5·8	15·3	1909	7·1	15·6
1902	6·0	15·3	1910	7·2	15·1
1903	6·2	15·4	1911	7·2	14·1
1904	6·3	15·4	1912	6·9	11·2
1905	6·6	16·1	1913	6·7	11·1
1906	6·8	16·0	1914	6·4	10·4
1907	6·8	15·6			

The coefficient of correlation in this case comes out negative and $= -\cdot 15$, but it is very small and probably not significant. If it were, it would imply that as indoor pauperism diminishes outdoor pauperism increases, and *vice versa.*

Example (4).—To find the correlation between the *Number of Cattle* and the *Number of Acres of Permanent Grass-land* in the *Coal-Producing Counties of England* (1915).

A Government Report was consulted giving the acreage under crops and grass and the number of live stock in each petty sessional division in the country, as returned on 4th June 1915, and the counties included were those which appear in the coal-mining reports published monthly in the *Labour Gazette.*

In each county the petty sessional divisions with the greatest and the least numbers of cattle and of acres of grass-land were noted, the numbers being written down to the nearest 1000, and, after a rough examination of the range of these variables from county to county, suitable class intervals were chosen and a table of double entry was drawn up, Table (29), with an empty square ready for each possible pair of variables.

TABLE (29). CORRELATION BETWEEN THE NUMBER OF CATTLE AND THE NUMBER OF ACRES OF PERMANENT GRASS-LAND IN THE COAL-PRODUCING COUNTIES OF ENGLAND (1915).

		Total Head of Cattle (expressed to nearest thousand)									
		x_1 0-5	x_2 5-10	x_3 10-15	15-20	20-25	25-30	30-35	x_p 35-40	*Totals*	*Mean x*
Total Number of Acres of Permanent Grass-land (expressed to nearest thousand)	y_1 0-5	10 15 150								15	2·50
	y_2 5-10	8 27 216	4 3 12							30	3·00
	y_3 10-15	6 30 180	3 18 54							48	4·37
	15-20	4 3 12	2 30 60							33	7·04
	20-25		1 25 25	0 5 0						30	8·33
	25-30	0 1 0	0 14 0	0 9 0	0 2 0					26	9·81
	30-35		-1 6 -6	0 22 0	1 3 3					31	12·02
	35-40		-2 1 -2	0 12 0	2 6 12	4 4 16				23	15·33
	40-45			0 3 0	3 4 12		9 1 9			8	16·87
	45-50			0 3 0	4 3 12	8 3 24		16 1 16		10	19·00
	50-55				5 4 20	10 4 40	15 1 15			9	20·83
	55-60				6 1 6	12 2 24		24 1 24	30 1 30	5	26·50
	60-65						21 1 21			1	27·50
	65-70						24 1 24			1	27·50
	70-75						27 1 27			1	27·50
	75-80							40 3 120		3	32·5
	y_q 80-85				11 1 11	22 1 22				2	20·0
	Totals	76	97	54	24	14	5	5	1	276	
	Mean y	9·14	20·13	33·24	42·33	50·00	59·50	67·50	57·5		

Each petty sessional division was then considered in turn and a dot was inserted in the particular square applicable to it : *e.g.* a petty sessional division with 42,000 acres of grass-land and feeding 19,000 cattle would be represented by a dot in the square defined by row (40-45) and col. (15-20) in Table (29) ; x was used to represent the number of cattle and y the number of acres of grass-land in any division, each expressed to the nearest 1000 units. All the dots were ultimately added in each square giving the frequency for each corresponding pair of variables, and these frequencies were recorded in the centres of the squares to which they applied : *e.g.* the frequency of petty sessional divisions stocking 10 to 15 thousand cattle and with 30 to 35 thousand acres under permanent grass was 22. The total frequency for each row, *i.e.* each array of selected y type, was also noted, in the column at the end of the rows : *e.g.* altogether 31 petty sessional divisions were observed of the type having 30 to 35 thousand acres of land under permanent grass. Likewise the total frequency for each column, *i.e.* each array of selected x type, was noted in the row at the foot of the columns : *e.g.* altogether 54 divisions were observed of the type stocking 10 to 15 thousand head of cattle.

It was possible now to treat each column separately and to calculate the mean y's associated with different types of x, namely x_1, x_2, x_3, . . . , and the frequencies so obtained were inserted in the bottom row of Table (29) : *e.g.* when x lies between 20 and 25 thousand, the mean value of y is 50 thousand. The resulting points—$(x_1, \bar{y}_1)$, $(x_2, \bar{y}_2)$, $(x_3, \bar{y}_3)$ . . . in the notation of Chapter x.—are plotted together in fig. (25), and they are seen to lie approximately in a straight line. The successive rows were treated in precisely the same way and the mean x's calculated corresponding to y's of different types, namely y_1, y_2, y_3, . . . , the frequencies obtained being recorded in the extreme right-hand column of Table (29) : *e.g.* when y lies between 45 and 50 thousand, the mean value of x is 19 thousand. The resulting points $(\bar{x}_1, y_1)$, $(\bar{x}_2, y_2)$, $(\bar{x}_3, y_3)$, . . . , are also plotted in fig. (25), and, excepting for values which depend upon only one or two records, they too lie roughly in a straight line which is not far from coinciding with the previous one, so that we shall expect on calculation to get a high value for the coefficient of correlation.

In order to calculate r we need first to find the mean and standard deviation for each variable. For this let us take as origin the point (12·5, 27·5). The essential details are shown immediately below the relative Tables (30) and (31).

TABLE (30). DISTRIBUTION OF PETTY SESSIONAL DIVISIONS ACCORDING TO THE HEAD OF CATTLE (EXPRESSED TO NEAREST 1000) STOCKED.

(1) No. of Cattle stocked (in thousands).	(2) Deviation from 12·5.	(3) No. of Petty Sessional Divisions.	(4) Product of Nos. in Cols. (2) & (3).	(5) Product of Nos. in Cols. (2) & (4).
	(x)			
0–5	−2	76	−152	304
5–10	−1	97	− 97	97
10–15	0	54	..	..
15–20	+1	24	+ 24	24
20–25	+2	14	+ 28	56
25–30	+3	5	+ 15	45
30–35	+4	5	+ 20	80
35–40	+5	1	+ 5	25
		276	−157	631

Mean number of cattle$=12{\cdot}5-\frac{157}{276}\times 5=9{\cdot}66$, since $\bar{x}=-\frac{157}{276}$ class units referred to 12·5 as origin; and $\sigma_x=5\sqrt{[\frac{631}{276}-(\frac{157}{276})^2]}$ $=5\sqrt{1{\cdot}963}=7{\cdot}00$.

[The numbers in col. (4) may be spoken of as the *first moments* of the totals of x arrays and the numbers in col. (5) as the *second moments*.]

In order to calculate easily the product deviation with reference to (12·5, 27·5) as origin, the value proper to each square was inserted just above the frequency and the product of the deviation by the frequency was inserted just below the frequency in different type of print to prevent confusion: *e.g.* the row (50-55) is +5 class intervals distant from the row (25-30) containing the origin, and the column (20-25) is +2 class intervals distant from the column (10-15) containing the origin; hence, for the particular square defined by this row and this column, the product deviation$=5\times 2=10$; also the frequency recorded in this square$=4$, so that it supplies a term 10×4 to the product deviation; the numbers 10, 4, and 40 are therefore the numbers which appear in the square. It is necessary to be careful with the signs; if the product deviation is to be positive, the separate deviations must be of like sign, both positive or both negative: hence they must either be both above or both below the numbers 12·5 and 27·5 respectively from which

they are measured. In this instance there are only two negative terms among the product deviations in the whole table.

TABLE (31). DISTRIBUTION OF PETTY SESSIONAL DIVISIONS ACCORDING TO THE NUMBER OF ACRES OF LAND (EXPRESSED TO NEAREST 1000) UNDER PERMANENT GRASS.

(1) No. of Acres under Grass (in thousands).	(2) Deviation from 27·5.	(3) No. of Petty Sessional Divisions.	(4) Product of Nos. in Cols. (2) & (3).	(5) Product of Nos. in Cols. (2) & (4).
	(y)			
0– 5	− 5	15	− 75	375
5–10	− 4	30	−120	480
10–15	− 3	48	−144	432
15–20	− 2	33	− 66	132
20–25	− 1	30	− 30	30
25–30	..	26	..	..
30–35	+ 1	31	+ 31	31
35–40	+ 2	23	+ 46	92
40–45	+ 3	8	+ 24	72
45–50	+ 4	10	+ 40	160
50–55	+ 5	9	+ 45	225
55–60	+ 6	5	+ 30	180
60–65	+ 7	1	+ 7	49
65–70	+ 8	1	+ 8	64
70–75	+ 9	1	+ 9	81
75–80	+10	3	+ 30	300
80–85	+11	2	+ 22	242
		276	−143	2945

Mean number of acres $=27{\cdot}5-\frac{143}{276}\times5=24{\cdot}91$, since $\bar{y}=-\frac{143}{276}$ class units ; and $\sigma_y=5\sqrt{[\frac{2945}{276}-(\frac{143}{276})^2]}=5\sqrt{10{\cdot}402}=16{\cdot}12$.

[The numbers in col. (4) are the *first moments* of the totals of y arrays, and the numbers in col. (5) are the *second moments*.]

It is now a simple matter to sum the product deviation terms, taking each column (or each row) in turn : *e.g.* the first column gives

$$150+216+180+12=558\,;$$

the second column gives

$$12+54+60+25-6-2=143,$$

and so on ; and, summing these results together, we get

$$558+143+76+126+96+160+30=1189.$$

But this is the sum of all the product deviations referred to (12·5, 27·5) as origin. Transferring now to the mean, we have

$$p=\tfrac{1189}{276}-\bar{x}\bar{y}$$
$$=\tfrac{1189}{276}-(-\tfrac{157}{276})(-\tfrac{143}{276})$$
$$=4{\cdot}013, \textit{ expressed in class units.}$$

Hence, $$r=p/\sigma_x\sigma_y,$$

where σ_x and σ_y are also to be expressed in class units,

$$=4{\cdot}013/\sqrt{(1{\cdot}963)}\sqrt{(10{\cdot}402)}$$
$$=0{\cdot}89,$$

a result not far from unity, so that the correlation is high.

The regression of 'acreage of grassland' (Y) on 'head of cattle' (X) is given by

$$(\mathrm{Y}-24{\cdot}91)=r\frac{\sigma_y}{\sigma_x}(\mathrm{X}-9{\cdot}66)$$
$$=(0{\cdot}89)\frac{(16{\cdot}12)}{(7{\cdot}00)}(\mathrm{X}-9{\cdot}66),$$

i.e. $$\mathrm{Y}=2{\cdot}05\mathrm{X}+5{\cdot}11.$$

The points representing the mean y's for x's of different types should lie close to this line which is shown in fig. (25). This equation enables us to predict the acreage under permanent grass to be found *on the average* in petty sessional divisions with a given total head of cattle in each. The words 'on the average,' to be tacitly understood even if not stated in all such cases, are emphasised because the prediction relates to the whole array of divisions of a particular type, and as it only professes to give the mean or most likely result it is not to be pronounced worthless if it fails in an individual trial with a selected division.

Again, the regression of X on Y is given by

$$(\mathrm{X}-9{\cdot}66)=r\frac{\sigma_x}{\sigma_y}(\mathrm{Y}-24{\cdot}91)$$

i.e. $$\mathrm{X}=0{\cdot}39\mathrm{Y}+0{\cdot}05,$$

which tells us the total head of cattle (X) to be found *on the average* in petty sessional divisions when the acreage under permanent grass (Y) is known. This line is also drawn in fig. (25).

Example (5).—The data for this example are taken from an exceedingly interesting Government Report on the Cost of Living of the Working Classes (*Report of an Inquiry by the Board of Trade into Working Class Rents and Retail Prices together with the Rates*

of Wages in certain Occupations in Industrial Towns of the United Kingdom in 1912 *in continuation of a similar Inquiry in* 1905.

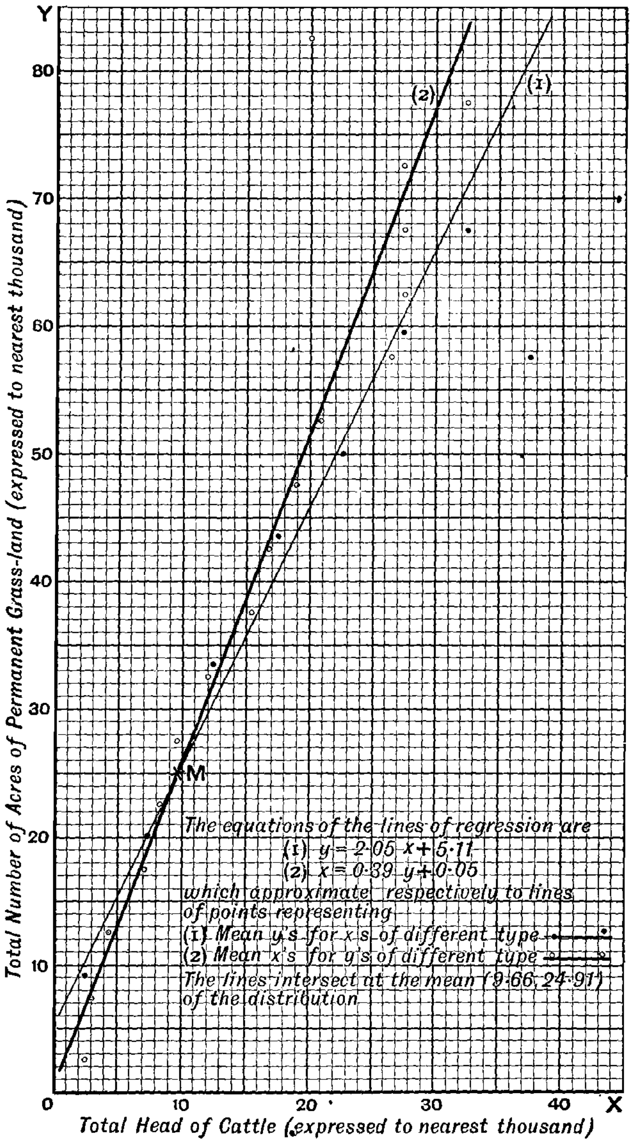


Fig. (25).

Cd. 6955). Some further particulars concerning this Report \ be found on p. 281.

The towns included in the inquiry numbered 93, but in five instances it was found desirable to consider closely adjacent municipalities as single towns thus reducing the number of town-units to 88, namely 72 in England, 10 in Scotland, and 6 in Ireland. In the example which follows the three zones of London, middle, inner, and outer, have been treated as separate towns, so making the net number of town-units 90. This number is too small to allow any real value to be attached to our results, but the fewness of the observations makes them easier to deal with as an illustration of method.

We begin as before by choosing convenient class intervals for the two factors we propose to consider, namely, *Increment of Unskilled Wages* and *Increment of Rents*—by increment in each case is meant the percentage increase (+) or decrease (−) between 1905 and 1912—and then form a correlation table. In the last example separate tables were drawn up to find means and S.D.'s, but that was only done in order to keep the argument clear at its first presentment : generally we may dispense with these additional tables and show all the working in one (see Table (32)).

The increment of wages runs from (−2·5) per cent. to (+11·5) per cent., so that, if we take (−0·5) as origin and a difference of 2 per cent. as unit, the classes run from (−1) to (+6), these numbers being shown in different type in the table, but in the same compartments as the others. In the fourth row from the bottom are shown the total frequencies for x arrays from class (−1) to class (+6), and in the row just below it these several frequencies are shown multiplied by their corresponding deviations measured from (−0·5) as origin in terms of the class unit—the resulting numbers give the first moments of the totals of x arrays. These numbers, multiplied again by their corresponding deviations, give the second moments of the totals of x arrays, and appear in the last row but one of the table.

We deal in exactly the same way with increment of rents : a percentage increment of (−1) is taken as origin from which deviations are measured, a difference of 3 per cent. is taken as unit, and the different classes then have deviations running from (−3) to (+6). The totals of y arrays, the first moments, and the second moments of these totals appear in the last three columns on the right-hand side of Table (32).

To calculate the deviation products, numbers were inserted in each square on the same principle as in the last example, and the sums of these products for each x array, that is for each column,

are given in the bottom row of the table—1, 0, 14, 6, etc., making in all a total of 126.

TABLE (32). CORRELATION BETWEEN INCREMENT OF UNSKILLED WAGES AND INCREMENT OF RENTS IN CERTAIN INDUSTRIAL TOWNS OF THE UNITED KINGDOM.

		x = Percentage Increment of Wages										
		-1 -2·5	0 -0·5	+1 1·5	+2 3·5	+3 5·5	+4 7·5	+5 9·5	+6 11·5	Totals of y arrays	1st. moments of y arrays	2nd. moments of y arrays
y = Percentage Increment of Working Class Rents: -3	-10		0 1 0							1	-3	9
-2	-7	2 1 2	0 3 0							4	-8	16
-1	-4		0 4 0	-1 1 -1	-2 2 -4		-4 1 -4	-5 1 -5	-6 1 -6	10	-10	10
0	-1		0 15 0	0 1 0	0 6 0	0 6 0		0 2 0		30	–	–
+1	2	-1 1 -1	0 9 0	1 1 1	2 3 6		4 3 12	5 1 5		18	18	18
+2	5		0 6 0	2 1 2	4 1 4		8 1 8	10 2 20		11	22	44
+3	8		0 3 0	3 4 12		9 1 9	12 1 12	15 2 30		11	33	99
+4	11		0 3 0							3	12	48
+5	14		0 1 0							1	5	25
+6	17						24 1 24			1	6	36
Totals of x arrays		2	45	8	12	7	7	8	1	90	75	305
1st. moments of x arrays		-2	–	8	24	21	28	40	6	125		
2nd. moments of x arrays		2	–	8	48	63	112	200	36	469		
Product Sums of x arrays		1	0	14	6	9	52	50	-6	126	Total Product Sum	

The necessary calculations are as follows :—

1. Mean $x = -0·5 + 2(125)/90 = 2·28$,

$$\sigma_x = 2\sqrt{[\tfrac{46}{90} - (\tfrac{125}{90})^2]} = 2\sqrt{(26585)}/90.$$

2. Mean $y = -1 + 3(75)/90 = 1·50$,

$$\sigma_y = 3\sqrt{[\tfrac{305}{90} - (\tfrac{75}{90})^2]} = 3\sqrt{(21825)}/90.$$

3. $p = \frac{126}{90} - (\frac{125}{90})(\frac{75}{90}) = \frac{1965}{(90)^2}$, *expressed in class units.*

Hence
$$r = p/\sigma_x\sigma_y$$
$$= \frac{1965}{(90)^2} \times \frac{90}{\sqrt{(26585)}} \times \frac{90}{\sqrt{(21825)}}$$
$$= 0{\cdot}08.$$

In substituting for σ_x and σ_y to find r we have omitted the factors 2 and 3 respectively, because the S.D.'s have to be expressed in the same units as p. Alternatively, if we worked with a difference of 1 per cent. as unit, instead of taking a difference of 2 per cent. as unit for x deviations, and a difference of 3 per cent. as unit for y deviations, each individual product of x and y deviations would

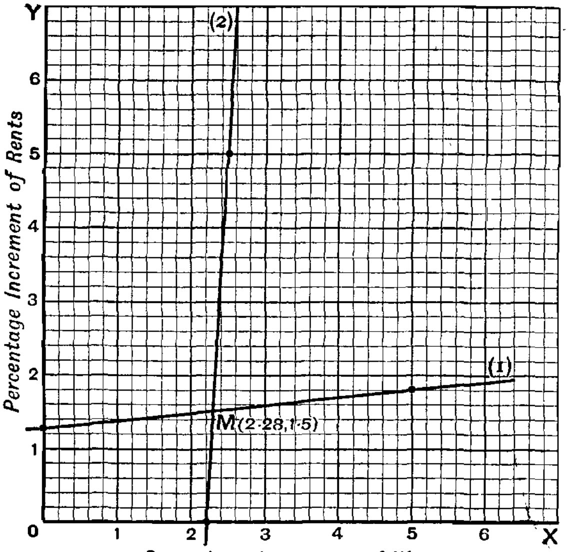


FIG. (26).

have to be multiplied by 2×3. Thus p would then be $6 \times 1965/(90)^2$, and we should get the same result for r as before by taking σ_x and σ_y as in (1) and (2) above. In this case r is so small as to be quite insignificant of any correlation between the two factors discussed, and the regression lines should therefore be not far from perpendicular to one another.

The regression of y on x, or the equation giving the most probable y for a given type x is

$$(y - 1{\cdot}50) = r\frac{\sigma_y}{\sigma_x}(x - 2{\cdot}28),$$

i.e.
$$y = 0{\cdot}11x + 1{\cdot}25.$$

Similarly, the regression of x on y is

$$x = 0{\cdot}06y + 2{\cdot}2.$$

To draw the first line we note that it passes through the points (0, 1·25) and (5, 1·8); also the second line goes through the points (2·2, 0) and (2·5, 5). The two lines intersect at M (2·28, 1·5), the mean of the distribution. They are drawn together in fig. (26).

TABLE (33). CORRELATION BETWEEN UNSKILLED WAGES AND RENTS IN CERTAIN INDUSTRIAL TOWNS OF THE UNITED KINGDOM.

y = *Index Number for Rents of Working-class Dwellings*	*x = Index Number for Wages of Unskilled Labour*									
	45·5	51·5	57·5	63·5	69·5	75·5	81·5	87·5	93·5	99·5
40·5	2				3			1		
48·5			2	1	3	1	4	3	1	
56·5		1		1		2	7	15	6	2
64·5					2	1	3	9	4	1
72·5				1			3	3	2	
80·5						1		1		1
88·5										1
96·5										1
104·5										
112·5										1

Example (6).—Instead of discussing the *Changes in Wages and Rents* between 1905 and 1912, it might be of interest to find the correlation between index numbers representing *Actual Wages and Rents* in October 1912, taken from the same Report. The necessary data for this purpose appear in Table (33) showing the distribution of frequency between the different classes: *e.g.* seven towns were observed in which the index number for wages was between the limits (79-84) and the index number for rents was between the limits (53-60). The wages figures quoted in Table (33) refer only to unskilled labour in the building trade; the inquiry actually embraced certain occupations in the building, engineering, and

printing trades, these having been selected as industries which are found in most industrial towns, and in which the time rates of wages are largely standardised.

TABLE (34). CORRELATION BETWEEN INCREMENT OF WORKING CLASS PRICES AND INCREMENT OF WORKING CLASS RENTS IN CERTAIN INDUSTRIAL TOWNS OF THE UNITED KINGDOM.

y = Percentage Increment of Rents	x = Percentage Increment of Prices						
	7·5	9·5	11·5	13·5	15·5	17·5	19·5
-10					1		
-7	1		1			2	
-4	1	2	2	2	1	2	
-1	1	4	6	10	8	1	
2		1	2	5	8	1	1
5	2		4	2	3		
8		1	2	3	1	4	
11				1	1		1
14				1			
17				1			

The coefficient of correlation turns out to be 0·46, distinctly larger than in the previous case. Also the lines of regression are :—

(1) $y = 0{\cdot}47x + 21$. (2) $x = 0{\cdot}45y + 56$.

Example (7).—The Report also furnishes data for evaluating the correlation between the *Increment of Working Class Prices and Increment of Working Class Rents*, again meaning by increment the percentage increase (+) or decrease (—) between 1905 and 1912 (see Table (34)).

The correlation in this case is very small, being only 0·13. The regression equations are :—

(1) $y = 0{\cdot}22x - 1{\cdot}5$. (2) $x = 0{\cdot}07y + 13$.

PART II

CHAPTER XII

INTRODUCTION TO PROBABILITY AND SAMPLING

SUPPOSE we wish to know the average measurement of some organ or character, *e.g.* length of forearm or weight or anything similar, in a large population containing several thousand individuals. The mean obtained by actual measurement if it were practicable to carry it out on so large a scale, would evidently depend to some extent upon the sex, the race, the age, the social class, and so on, of the individuals selected, and we shall accordingly assume our population to be composed of individuals of the same race and sex, at about the same age, taken from the same class, etc.; it would be impossible in practice no doubt to secure that all conditions should be identically the same for all the individuals observed, but the population may be as homogeneous as we care to make it in theory.

Now suppose that, instead of attempting to measure every single individual, a random sample of 1000 from among the population be taken and that the mean and variability of the measurements for this sample be calculated, giving results m_1 and σ_1. With these may be compared m_2 and σ_2, the results of measuring a second sample of 1000 individuals, m_3 and σ_3, the results of a third sample, and so on. It is extremely unlikely that the values obtained for the m's in this way will equal one another, neither will the σ's be equal; but, if we have succeeded at the beginning in avoiding all ill-balanced influences when we tried to make the field of observation as homogeneous as possible, the resulting m's and σ's will only differ from the values of the mean and variability for the whole population, assuming they could be measured, within a comparatively small range.

Differences of this kind, which arise merely owing to the fact that we are often obliged in practice, for lack of time or means, to deal with a comparatively small sample instead of with the whole population of which it forms a part, are said to be *due to random*

sampling. Granted that the samples themselves are adequate in size (containing, say, from 500 to 1000 individuals each) an estimate of differences to be expected between one and another can be made, and unless the observed differences fall outside recognized limits it is said that they are *not significant* of any difference other than such as might quite well be accounted for by random sampling alone.

In theory, then, we can imagine a large number of such random samples selected, and by determining the S.D. of their means, m_1, m_2, m_3, . . . , we should have a fair measure of the deviation which might quite well occur from the true value, that is, from the mean of the population as a whole, through working only with a sample. Further, a range of two or three times the S.D. on either side of the true mean ought to take in the majority of the sample means observed.

Exactly the same principle holds good in dealing with the proportion of individuals in a given population which can be assigned to a particular class, or in discussing the S.D. of the distribution, or the C. of V., or a coefficient of correlation, or any other statistical constant, no matter what the nature of the character may be which is measured or observed, or whether it relates to animate or inanimate objects. Take, for instance, the variability—by selecting several samples from a given population we get a series of values σ_1, σ_2, σ_3 . . ., and in the S.D. of this distribution of variabilities we have a measure to which we can compare the deviation of any sample variability, σ_r, from the true variability of the whole population, while a range two or three times the S.D. might be expected to include the majority of the different variabilities met with in the samples.

Although the S.D., as we have explained, provides quite a suitable measure of the extent of deviation of a sample constant from its true value in the population as a whole, in practice, owing to the historical development of the theory having followed the track of the normal curve of error [see Chapter XVIII.] a measure known as the *probable error* and equal roughly to two-thirds of the S.D. is not seldom employed in its place. The main, if not the sole, justification for retaining this measure is that it has established its position by long usage, and in any case it is very easily deduced from the S.D. by the relation

$$\text{p.e.} = 0{\cdot}6745\ \text{S.D.},$$

which follows at once from the normal curve and is only strictly

justified when the distribution is normal (see p. 246). Let it suffice here that instead of simply using the S.D., as might now seem the obvious course, some writers prefer to multiply the S.D. by a certain fraction, in which there is no particular virtue except that which arises through honourable descent, and to work with the 'probable error.'

Since we do not know how much weight to assign to any result unless the magnitude of its p.e. is also given, results are frequently stated in the following manner: in a study of the *Variation and Correlation in the Earthworm*, by R. Pearl and W. N. Fuller [*Biometrika*, vol. iv. pp. 213-229]:—

$$\text{Mean length of worm} = 19{\cdot}171 \pm 0{\cdot}094 \text{ cms.},$$
$$\text{S.D.} = 3{\cdot}077 \pm 0{\cdot}067 \text{ cms.},$$
$$\text{C. of V.} = 16{\cdot}049 \pm 0{\cdot}356 \text{ per cent.},$$

meaning that the mean length of the worms measured was 19·171 cms., subject to a probable error of 0·094 cms. which might be in excess or defect, in other words the mean length lay probably somewhere between

19·077 cms. and 19·265 cms.;

similar remarks apply to the variability, absolute (S.D.) or relative (C. of V.).

When the standard deviation (p.e./0·6745) is used as the measure of error due to simple sampling, the fact is generally recorded, and it is sometimes spoken of as the *standard error* in that connection, but, as it seems unnecessary to multiply names for ideas which are not really new, only that they appear in a new setting, we shall not employ the term.

It must be clearly understood that no outstanding and predictable cause exists, by our hypothesis, for such differences as occur in the statistical constants between one sample and another: they are the resultant effect of a complex of forces which cannot be properly traced, still less measured, apart from one another, and which have been happily described as that 'mass of floating causes generally known as chance.' Since therefore the forces coming into play, under the ideal conditions formulated, are of the same chance nature as those affecting the spin of a well-balanced coin or the selection of a card from a smooth and well-shuffled pack, it may be expected that the resulting distribution of means, $m_1, m_2, m_3, \ldots$, of S.D.'s, $\sigma_1, \sigma_2, \sigma_3, \ldots$, and of all the other constants will likewise be subject to the same laws of probability

which serve to describe within limits what happens in the case of coin or card. It follows that some acquaintance with the first elements of mathematical probability is essential if one is to understand the theory of sampling, and a short digression must here be made in order to introduce that subject. This will be found to lead directly to a solution, under certain prescribed conditions, in the simple case when the character observed is an attribute like complexion, fair or dark, or like birth, male or female, which can only fall into one of two definite classes and when every one observation in the sample is independent of every other. In the more general case where the character observed is capable of direct measurement and may lie in magnitude anywhere along a scale of values divided up into a number of different classes, it is not so easy to determine the effect of random sampling, because it is not possible, as it is in the previous case, actually to draw up a frequency table describing in detail the character of the distribution to be expected from theory in any given sample.

The idea contained in the word probability is one familiar to us in our everyday talk, but if we seek to analyse it as used we find it as elusive as the personality of the user. A remarks: 'Wars will probably be stamped out, like duelling, in the course of time.' B replies: 'No! fighting will probably go on as long as the world lasts—you can't change human nature.' Now the amount of credence we are prepared to give to each of these statements is vague and uncertain until we know something about A and B themselves and the value of their judgment, quite apart from the influence of our own opinion upon the matter; perhaps A is an optimist or B is a pessimist, and in estimating the 'probably' used by each we must allow for these facts. Probability, then, in ordinary conversation, is something largely subjective: it has a varying significance according to the person who uses the word and, unless we could get rid of this personal element, it would be hopeless to try and approach it along scientific lines.

Mathematical probability is unlike colloquial probability in that all the uncertainty is taken out of it, or at least the uncertainty is confined within defined limits. We shall only touch the fringe of the subject in this book, and what we have to say may be best introduced by considering some examples which may appear trivial, but they possess the merit that no personal bias can enter into their discussion to distort the results. The reader must not be impatient at their artificial character: in many, if not in all, branches of science, before tackling any particular problem as it

actually exists, it is helpful to examine what can be deduced in a simple case free from all complication, and, having settled that, we try to see how the results are affected when we come to allow one by one for the various complicating factors which exist. For example, in Astronomy, the track of a planet in space may first be found on the hypothesis that the sun alone is the compelling influence. Then we may proceed to discuss how it is deflected from its path when the gravitational influence of neighbouring planets also is taken into account.

Let us start with an ordinary pack of playing cards, and, after shuffling, turn up one card. Can we measure the probability that this card shall be (1) the 7 of spades? (2) some spade?

Altogether there are 52 cards, and we will suppose that the cards are so cut and so smooth that each of the 52 has an equal chance of being turned up: for instance, there is to be no stickiness or anything to help any particular card to evade us by sticking fast to its neighbour. Now we are *certain* to turn up *some* card and there are 52 different possibilities, each of them *by hypothesis* equally probable. If, then, we agree to denote certainty by unity, we must divide 1 into 52 equal parts and assign one part to each card as the probability of its appearance.

1. The probability (or *chance* as it is sometimes called) of turning up any stated card, such as the 7 of spades, is therefore 1 out of 52, *i.e.* 1/52.

2. Again, since there are 13 spades in all, the chance of turning up some spade is 13 out of 52, *i.e.* $13/52 = 1/4$.

These results may be put in another way which is often useful. If the experiment is repeated a great number of times, a return to the initial conditions of the problem being made after each trial by replacing the card drawn and reshuffling the pack, we should expect to turn up the 7 of spades *on the average* about once in every 52 experiments, and we should expect to turn up some spade *on the average* about once in every 4 experiments. This must not be taken to mean that in 4 experiments we are sure to turn up just one spade—a trial will readily prove such a statement to be untrue—but that, if we went on performing experiment after experiment, we should *in the long run* get a proportion of about 1 spade to every 4 experiments and a trial will likewise prove the truth of this statement.

Generally, when an event can happen in n different ways altogether, and among these different ways there are a which give what might be called successful events, the probability of success

at any single happening is a out of n, *i.e.* a/n, and is usually denoted by the letter p, and the probability of failure is $(n-a)$ out of n, *i.e.* $(n-a)/n$, and is usually denoted by the letter q.

Clearly $(p+q)=1$, and this is reasonable because we are certain to get either a success or a failure at a single trial and unity was fixed as the measure of certainty. In k trials, the probable number of successes would be kp and of failures kq, because in n trials, on the average, there are a, or np, successes and $(n-a)$, or nq, failures.

Example (1).—In the second case considered above, the probability of success (turning up a spade) is a out of n

$$=a/n=13/52=1/4=p,$$

and the probability of failure (not turning up a spade, *i.e.* turning up one of 39 other cards) is $(n-a)$ out of n

$$=(n-a)/n=39/52=3/4=q.$$

And $$(p+q)=1/4+3/4=1.$$

Example (2).—What is the chance of drawing either a picture card or an ace from the pack at a single trial ?

Altogether there are 12 picture cards, and the chance of drawing any one of them is thus 12 out of 52

$$=12/52=3/13\,;$$

and the chance of drawing any one of the 4 aces is 4 out of 52

$$=4/52=1/13.$$

Hence the total probability required

$$=3/13+1/13=4/13.$$

Generally, if the probability of one type of event is p_1, and the probability of a second type of event is p_2, and if either type is reckoned a success, then the total probability of success is (p_1+p_2). This evidently holds good however many different types there may be, and even if there is only one event of each type.

Consider now the simultaneous happening of two events, one of which can happen in n different ways, a among which are to be regarded as successful, and the second can happen in n' different ways, a' among which are to be regarded as successful. Further, the two events are to be absolutely independent of one another in the sense that neither is to influence the success or failure of the other. What is the probability of a double success occurring ?

The total number of different combinations of the two events

possible is nn', for any one of the n possible happenings for the first event can be combined with any one of the n' possible happenings for the second event. Also the total number of different combinations of two successes possible is aa', for any one of the a possible successes for the first event can be combined with any one of the a' possible successes for the second event. Hence, according to our definition of probability, the probability of a double success is aa' out of $nn' = aa'/nn' = (a/n)(a'/n')$.

Thus to get the probability of a double success for a combination of two independent events we must multiply together the separate probabilities for the success of each event taken by itself.

Similarly, in the above case, the probability of a double failure $=(n-a)(n'-a')/nn'$; and the probability of one success and one failure

$$=\frac{a}{n}\cdot\frac{n'-a'}{n'}+\frac{n-a}{n}\cdot\frac{a'}{n'},$$

for the first event can be a success and the second a failure or the first a failure and the second a success.

Here, again, if we take all the different possibilities into account, and add the probabilities corresponding to each case, we arrive at certainty, the measure of which is unity, thus :—

probability of 2 successes $=aa'/nn'$,

,, 1 success and 1 failure $=a(n'-a')/nn'+a'(n-a)/nn'$

,, 2 failures $=(n-a)(n'-a')/nn'$.

Therefore total probability, all cases,

$$=\frac{aa'}{nn'}+\frac{a(n'-a')}{nn'}+\frac{a'(n-a)}{nn'}+\frac{(n-a)(n'-a')}{nn'}$$

$$=(aa'+an'-aa'+a'n-a'a+nn'-na'-an'+aa')/nn'$$

$$=nn'/nn'$$

$$=1.$$

Example.—Take two packs of cards. What is the probability of drawing an ace from the first pack and a king, queen, or knave from the second pack ?

Here $a=4$, $n=52$, $a'=12$, $n'=52$; hence the required probability

$$=aa'/nn'=4/52\times 12/52=3/169=1/56\tfrac{1}{3}.$$

Thus we might expect to succeed on the average about once in 56 trials.

We proceed to discuss the case of a coin spun a number of times in succession, and we shall find the probabilities of the appearance of so many heads (H) and so many tails (T) in so many spins on the hypothesis that the coin is perfectly balanced and equally likely to fall on either side.

In 1 *spin* there are 2 possible events, namely H or T, which we shall write simply as

$$(H, T).$$

In 2 *spins* there are 4 possible events, because we can combine the H or T of the first with an H or T at the second spin, and we may express the result thus

$$(H, T)(H, T)=(HH, HT, TH, TT) ;$$

the interpretation of which is that we may get either head followed by head, or head followed by tail, or tail followed by head, or tail followed by tail.

In 3 *spins* there are 8 possible events, because we can combine the 4 events previously possible with an H or T at the third spin, thus getting

$$\begin{aligned}&(H, T)(H, T)(H, T)\\&\quad=(H, T)(HH, HT, TH, TT)\\&\quad=(HHH, HHT, HTH, HTT, THH, THT, TTH, TTT) ;\end{aligned}$$

the interpretation of which is that we may get either 3 heads in succession, or 2 heads followed by 1 tail, or head followed by tail followed by head, and so on.

In 4 *spins* there are 16 possible events, because we can combine the 8 events previously possible with an H or T at the fourth spin, thus

$$\begin{aligned}&(H, T)(HHH, HHT, HTH, HTT, THH, THT, TTH, TTT)\\&\quad=(HHHH, HHHT, HHTH, HHTT, HTHH, HTHT,\\&\qquad HTTH, HTTT, THHH, THHT, THTH, THTT,\\&\qquad TTHH, TTHT, TTTH, TTTT).\end{aligned}$$

But the method here adopted to get the possible events at each stage is precisely the same as that which gives the successive terms in the ordinary algebraical expansions of

$$(H+T),\ (H+T)(H+T),\ (H+T)(H+T)(H+T),\ \text{etc.}$$

Also each new spin has the effect of doubling the number of possible

events obtained at the previous spin, and we conclude **that** ***in n spins***, there are

$$(2\times 2\times 2\times \ . \ . \ . \text{ to } n \text{ factors}),$$

or 2^n, possible events, and these events are given by the successive terms in the expansion of

$$[(H+T)(H+T)(H+T) \ . \ . \ . \text{ to } n \text{ factors.}]$$

Let us now consider the probabilities of the different events obtainable. The important point to notice is that at any stage each possible event has exactly the same probability, for there is no reason why any particular spin should give H rather than T, or T rather than H: for example, in 3 spins there are 8 possible events, each by itself equally probable, and we therefore divide the unity of certainty into 8 equal parts and assign one part to each event, thus

$$\begin{aligned}
&\text{probability of 3 heads—HHH}=\tfrac{1}{8}\\
&\text{probability of 2 heads and 1 tail—}\left.\begin{matrix}\text{HHT}=\tfrac{1}{8}\\ \text{HTH}=\tfrac{1}{8}\\ \text{THH}=\tfrac{1}{8}\end{matrix}\right\}\tfrac{3}{8}\\
&\text{probability of 1 head and 2 tails—}\left.\begin{matrix}\text{HTT}=\tfrac{1}{8}\\ \text{THT}=\tfrac{1}{8}\\ \text{TTH}=\tfrac{1}{8}\end{matrix}\right\}\tfrac{3}{8}\\
&\text{probability of 3 tails—TTT}=\tfrac{1}{8}.
\end{aligned}$$

It is clear from this arrangement that, *if the order of the appearance of H and T is indifferent,* some events are of the same type and some types are likely to appear oftener than others, *e.g.* the probability of getting '2 heads and 1 tail' (or '1 head and 2 tails') is three times as great as the probability of getting '3 heads' or '3 tails.' Hence for conciseness it is convenient to adopt the ordinary index notation and write

$$HHH=H^3,\ HHT=H^2T,\ HTH=H^2T, \text{ etc.},$$

so that the possible events in 3 spins are

$$H^3,\ 3H^2T,\ 3HT^2,\ T^3;$$

in 4 spins they are

$$H^4,\ 4H^3T,\ 6H^2T^2,\ 4HT^3,\ T^4;$$

and so on.

The probability of any particular type is now readily written down: *e.g.* in 4 spins, the probability of getting 2 heads and 2 tails

=(number of successful events possible)/(total number of events possible)

$=6/2^4=6/16=\tfrac{3}{8}$.

But the binomial expansion always sums together terms of the same type for us in just the manner wanted, and we have the possible events in n spins given by the successive terms in the expansion of

$$(H+T)(H+T)(H+T) \dots \text{ to } n \text{ factors,}$$

i.e. $$(H+T)^n,$$

i.e. $$H^n + {}^nC_1 \cdot H^{n-1}T^1 + {}^nC_2 H^{n-2}T^2 + \dots + T^n,$$

and therefore again the probability of any particular combination is readily written down : *e.g.* probability of '$(n-2)$ heads, 2 tails'

$=$(number of successful events possible)/(total number of events possible)

$= {}^nC_2/2^n$.

Another way of stating the result obtained is to say that we might expect to get

n heads	appearing on the average	about	once	in every	2^n	trials,
$(n-1)$ heads, 1 tail	,,	,,	,,	${}^nC_1$ times	,,	,,
$(n-2)$ heads, 2 tails	,,	,,	,,	${}^nC_2$ times	,,	,,

and so on.

If, in accord with our previous notation, we call the appearance of, say, H at any spin a 'success,' and label its probability $\frac{1}{2}$ by the letter p, and if consequently the appearance of T at any spin is a 'failure,' its probability, $\frac{1}{2}$, to be labelled by the letter q, we have the probabilities of the different combinations of events in $(H+T)^n$, or

$$H^n + {}^nC_1 H^{n-1}T^1 + {}^nC_2 H^{n-2}T^2 + \dots + T^n,$$

given by the corresponding terms in $(p+q)^n$, or

$$p^n + {}^nC_1 p^{n-1}q^1 + {}^nC_2 p^{n-2}q^2 + \dots + q^n,$$

where $p=q=\frac{1}{2}$.

After each spin of the coin in the case considered the distribution of probabilities was symmetrical, *e.g.* after the fourth spin the probabilities were

$$\tfrac{1}{16}, \tfrac{4}{16}, \tfrac{6}{16}, \tfrac{4}{16}, \tfrac{1}{16}.$$

We pass on now to a case where the distribution is not symmetrical, owing to the fact that p and q are no longer equal for any isolated event.

Consider the throw of an ordinary die in which each of the six faces is assumed to have an equal chance of appearing uppermost. The probability of throwing, say, a 3 is 1/6, since we are *certain* to throw either 1, 2, 3, 4, 5, or 6 ; and the probability of failing to throw a 3 is 5/6, since we are *certain* either to throw a 3 or not to throw a 3.

If we represent the probability of success (say, in this case, throwing a 3) by p (*i.e.* 1/6), and failure (*i.e.* in this case, failing to throw a 3) by q (*i.e.* 5/6), we have

$$p+q=1/6+5/6=1.$$

Bearing in mind then that the probability for a combination of two independent events is determined by multiplying together the separate probabilities for each, we have the following table showing what might be expected when 1, 2, or 3 dice are thrown up together, where s stands for success and f for failure :—

No. of Dice thrown.	Different Possibilities.	Different Probabilities.
1	s, f.	p, q.
2	ss, sf.	pp, pq.
	fs, ff.	qp, qq.
3	sss, ssf, sfs, sff,	ppp, ppq, pqp, pqq,
	fss, fsf, ffs, fff.	qpp, qpq, qqp, qqq.

The table is easily extended on the same principle, and at each step, it will be noticed, a fresh pair of possibilities, s or f, is introduced, with corresponding p or q, to be combined with what has gone before.

If the order of appearance of s and f is a matter of indifference, *e.g.* if it does not matter whether the first die shows s and the second f, or *vice versa*, so that results of the type sff and fsf may be regarded as equivalent, we may use the index notation, as in the coin case, to render the table more concise, thus :—

No. of Dice thrown.	Different Possibilities.	Corresponding Probabilities.
1	s, f.	p, q.
2	s^2, $2sf$, f^2.	p^2, $2pq$, q^2.
3	s^3, $3s^2f$, $3sf^2$, f^3.	p^3, $3p^2q$, $3pq^2$, q^3.

When, therefore, n dice are thrown we again recognize the different possibilities as given by the successive terms in the expansion of $(s+f)^n$, namely

$$s^n+{}^nC_1s^{n-1}f^1+{}^nC_2s^{n-2}f^2+\ .\ .\ .\ +f^n,$$

and the corresponding probabilities by the successive terms in the expansion of $(p+q)^n$, namely

$$p^n+{}^nC_1p^{n-1}q^1+{}^nC_2p^{n-2}q^2+\ .\ .\ .\ +q^n.$$

Hence the probability of throwing n threes $=p^n=1/6^n$;

„ „ „ $(n-1)$ „ $={}^nC_1p^{n-1}q^1$

$$=n\cdot\frac{1}{6^{n-1}}\cdot\frac{5}{6}$$

$$=5n/6^n;$$

„ „ „ $(n-2)$ „ $={}^nC_2p^{n-2}q^2$

$$=\frac{n(n-1)}{1\cdot 2}\cdot\frac{1}{6^{n-2}}\cdot\frac{5^2}{6^2}$$

$$=25n(n-1)/2\,.\,6^n;$$

and so on.

The result we have just obtained is of perfectly general application. Whether we spin n coins, in which the probability, p, of success (say 'heads') for each is 1/2, or throw n dice, in which the probability, p, of success (say 'to get a 3') for each is 1/6, or have any n similar but independent events happening in which the probability of success for each is p, the different resulting possibilities as to success are given by the successive terms in the expansion of $(s+f)^n$, and their corresponding probabilities are given by the successive terms in the expansion of $(p+q)^n$.

We are thus in a position to form a frequency table, like that on p. 53, showing the probabilities of getting 0, 1, 2 . . . n successes (in other words, the proportional frequencies of these different numbers of successes) at the occurrence of n similar independent events, where p is the probability of success for each and q is the probability of failure :—

TABLE (35). BINOMIAL DISTRIBUTION.

(1) Number of Successes.	(2) Frequency.	(3) Product of Nos. in Cols. (1) & (2).	(4) Product of Nos. in Cols. (1) & (3).
(x)	(f)	(fx)	(fx^2)
0	q^n	0	0
1	$nq^{n-1}p^1$	$nq^{n-1}p^1$	$nq^{n-1}p^1$
2	$\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2$	$n(n-1)q^{n-2}p^2$	$2n(n-1)q^{n-2}p^2$
3	$\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}q^{n-3}p^3$	$\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$	$\frac{3n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$
.	. .	. .	. .
.	. .	. .	. .
n	p^n	np^n	n^2p^n
	1	np	$np[1+p(n-1)]$

Col. (1) gives the deviations from the origin of measurement, which in this case is taken as 'no successes,' the class interval being equal to a difference of 1 in the number of successes.

The summations of the last three columns are effected as follows :—

Col. (2). $q^n+nq^{n-1}p^1+\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2+\ \ldots\ +p^n$

$=(q+p)^n$

$=1,$

because $p+q=1.$

Col. (3).

$$nq^{n-1}p^1+n(n-1)q^{n-2}p^2+\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3+\ \ldots\ +np^n$$

$$=np\left[q^{n-1}+(n-1)q^{n-2}p^1+\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\ \ldots\ +p^{n-1}\right]$$

$$=np(q+p)^{n-1}$$

$$=np.$$

Col. (4).

$$nq^{n-1}p^1+2n(n-1)q^{n-2}p^2+\frac{3n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3+\ \ldots\ +n^2p^n$$

$$=np\left[q^{n-1}+2(n-1)q^{n-2}p^1+\frac{3(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\ \ldots\ +np^{n-1}\right]$$

$$=np\left[\left\{q^{n-1}+(n-1)q^{n-2}p^1+\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\ \ldots\ +p^{n-1}\right\}\right.$$

$$\left.+\left\{(n-1)q^{n-2}p^1+\frac{2(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\ \ldots\ +(n-1)p^{n-1}\right\}\right]$$

$$=np[(q+p)^{n-1}+(n-1)p\{q^{n-2}+(n-2)q^{n-3}p+\ \ldots\ +p^{n-2}\}]$$

$$=np[1+(n-1)p(q+p)^{n-2}]$$

$$=np[1+p(n-1)].$$

The arithmetic mean of the distribution

=sum of terms in col. (3)/sum of terms in col. (2)

$=\Sigma(fx)/\Sigma(f)$

$=np.$

The mean-square deviation referred to zero as origin, zero in this case corresponding to ' no successes '

$$=\text{sum of terms in col. (4)/sum of terms in col. (2)}$$
$$=\Sigma(fx^2)/\Sigma(f)$$
$$=np[1+p(n-1)].$$

Thus the **standard deviation**, σ, is given by

$$\sigma^2=np[1+p(n-1)]-\bar{x}^2,$$

where $\bar{x}$ is the deviation of the mean from the origin of **measurement**, so that $\bar{x}=np$.

Therefore
$$\sigma^2=np[1+p(n-1)]-n^2p^2$$
$$=np(1-p)+n^2p^2-n^2p^2$$
$$=npq.$$

Hence
$$\sigma=\sqrt{(npq)},$$

and
$$\text{p.e.}=0\cdot6745\sqrt{(npq)}.$$

These two results are exceedingly important, and it is essential to understand what it is they measure. An example may help to make this clear.

If we spin 300 coins, counting ' head ' for each a success, the number of heads we shall get will be unlikely to differ very greatly from the average or mean number of successes, np, *i.e.* 150 if $p=1/2$ for each coin, and in the long run, if we repeat the experiment a great number of times, we shall get a proportion of about 150 heads to every one experiment. Again, if we throw 300 dice, counting every throw of the number 5, say, for each die a success, so that p in this case$=1/6$, the number of fives we shall get will be unlikely to differ much from np, *i.e.* 50, and in the long run, if we repeat the experiment a great number of times, we shall get on the average a proportion of about 50 fives to every experiment; we should find, for example, something like 5000 fives if we threw 300 dice 100 times in succession. The arithmetic mean of the distribution tells us therefore about what number of successes to expect in one experiment with n events if n is fairly large, though we should be unlikely to get exactly this number if we confined ourselves to the one experiment.

The second result, the S.D., supplies us with a measure of the unlikelihood of getting the exact number of successes expected at any single experiment, for it defines the dispersion of the different numbers of possible successes about their average. Clearly the greater the dispersion, the greater is the likelihood of missing the

average. The mean number of successes when an experiment is repeated a great number of times is np, but at any single experiment it is not unlikely that the number of successes obtained may differ from np by as much as $0{\cdot}6745\ \sqrt{(npq)}$ in excess or in defect; it is, however, unlikely, as we shall see later (p. 244), that the number will differ from np by more than $3\sqrt{(npq)}$ in excess or defect when the distribution is not very skew, or unsymmetrical, especially if n be large. The probable error in the case above when we throw a sample of 300 dice is

$$=0{\cdot}6745\sqrt{(300\times 1/6\times 5/6)}=0{\cdot}6745\sqrt{(41{\cdot}67)}=4{\cdot}4,$$

and it is therefore quite likely that the number of fives obtained at one experiment will differ from the expected number, 50, by as much as 4 or 5 in excess or defect, but it is unlikely that the number will fall outside the limits $50\pm 3\sqrt{(41{\cdot}67)}$, say 30 to 70.

It is sometimes more convenient to refer to the *proportion* of successes, etc., expected at any experiment rather than to the actual number expected. In that case, since with n events the expected number of successes is pn, but the number obtained may quite likely differ from this by $\pm 0{\cdot}6745\sqrt{(npq)}$, therefore with n events the expected proportion of successes is pn/n, *i.e.* p, with quite possibly an error$=\pm 0{\cdot}6745\sqrt{(npq)}/n$, *i.e.* $\pm 0{\cdot}6745\sqrt{(pq/n)}$.

Thus, with the 300 dice, the expected proportion of successes at one experiment lies between

$$[1/6-0{\cdot}6745\sqrt{(1/6\times 5/6\div 300)}] \text{ and } [1/6+0{\cdot}6745\sqrt{(1/6\times 5/6\div 300)}]$$

i.e. $(1/6-0{\cdot}6745/46{\cdot}5)$ and $(1/6+0{\cdot}6745/46{\cdot}5)$

i.e. $1/5{\cdot}5$ and $1/6{\cdot}6$;

and it is unlikely that the proportion will differ from 1/6 by more than $3/46{\cdot}5$, *i.e.* $1/15{\cdot}5$.

To illustrate how the binomial distribution might be directly applied, an experiment was made with 900 digits selected at random by taking in succession the digits in the seventh decimal place in the logarithms of the following numbers :—

10054, 10154, 10254, . . . 99954,

as given in Chambers's Mathematical Tables. In this way each of the 10 digits, 0, 1, 2, 3 . . . 9, may be supposed to have stood an equal chance of selection each time one was written down. Gaps of 100 were left between the numbers selected so as to avoid runs

of the same figure which sometimes occur even in the seventh decimal place owing to lack of independence.

The digits were arranged in 36 columns, each column containing 25 digits, and in this way we obtained what was equivalent to 36 separate but like experiments with 25 events each. If we agree to regard the appearance of a 7 or an 8 as a successful event, and the appearance of any other digit as a failure, the chance of success at any appearance is 2/10, and the chance of failure is 8/10. The case is thus of exactly the same kind as that of throwing 25 dice 36 times in succession, and if the probability of success, namely 1/5, for each independent event, be denoted by p, and the probability of failure, namely 4/5, by q, the distribution of successes and failures should approximately conform to that given by the expansion of

$$(s+f)^{25}$$

for any particular experiment, and since the experiment was repeated 36 times, the total numbers of successes and failures of different orders obtained should approximately conform to

$$36(p+q)^{25},$$

for if the probability of an event is p the number of events to be expected in N trials is Np.

The actual distribution observed is compared with that given by the binomial expansion in Table (36). Col. (2) is obtained by picking out the appropriate terms in the expansion of $36(p+q)^{25}$, where $p=1/5$, $q=4/5$; this expansion is

$$36\left(p^{25}+\frac{25}{1}\cdot p^{24}q^{1}+\frac{25\cdot 24}{1\cdot 2}p^{23}q^{2}+\ \ldots\ +q^{25}\right).$$

Thus, 5 successes occur

$$36\frac{25\cdot 24\ .\ .\ .\ 6}{1\cdot 2\cdot 3\ .\ .\ .\ 20}p^{5}q^{20}$$

times, and this equals 7·06, or approximately 7.

The mean number of successes by theory$=np=25/5=5$. The mean by trial, since it is measured from zero as origin, the numbers in col. (1) being the deviations,

$$=\Sigma(fx)/\Sigma(f)=162/36=4{\cdot}5.$$

The standard deviation by theory

$$=\sqrt{(npq)}=\sqrt{(25\times\tfrac{1}{5}\times\tfrac{4}{5})}=2.$$

TABLE (36). DISTRIBUTION OF SUCCESSES (GETTING A 7 OR 8) IN THE RANDOM CHOICE OF 25 DIGITS 36 TIMES IN SUCCESSION.

(1) No. of Successes.	(2) Frequency by Calculation.	(3) Frequency by Experiment.	(4) Product of Nos. in Cols. (1) & (3).	(5) Product of Nos. in Cols. (1) & (4).
(x)		(f)	(fx)	(fx^2)
1	1	1	1	1
2	3	5	10	20
3	5	5	15	45
4	7	7	28	112
5	7	9	45	225
6	6	4	24	144
7	4	3	21	147
8	2	0	0	0
9	1	2	18	162
	36	36	162	856

By trial, the mean square deviation, measured from zero as origin

$$=\Sigma fx^2/\Sigma f$$
$$=856/36.$$

Thus the S.D. by trial $=\sqrt{(\frac{856}{36}-\bar{x}^2)}$,

where $\bar{x}$ is the deviation of the mean from the origin,

$$=\sqrt{[856/36-(4\cdot5)^2}$$
$$=1\cdot88.$$

It will be seen that not one of the 36 experiments gave a number of successes differing from 5, the theoretical mean, by more than twice the S.D., for the number ranges only between 1 and 9.

If we treat the 900 digits as 900 separate experiments with one event each, instead of treating them as 36 experiments containing 25 events each, we have 1/10 as the chance for the appearance of any particular digit, and hence the number of times any digit may be expected to appear

$$=np\pm\tfrac{2}{3}\sqrt{(npq)}, \text{ approximately}$$
$$=(900)\tfrac{1}{10}\pm\tfrac{2}{3}\sqrt{(900\times\tfrac{1}{10}\times\tfrac{9}{10})}$$
$$=90\pm6.$$

The actual number of occurrences of each digit was as follows :—

Digit	0	1	2	3	4	5	6	7	8	9
No. of Occurrences .	95	96	93	105	91	80	82	72	90	96

so that the digit 7 showed the greatest divergence from 90 of any, and this was only just three times the probable error.

[The Theory of Probability is older than that of Statistics. Todhunter, in his *History*, states that 'writers on the subject have shown a justifiable pride in connecting its true origin with the great name of Pascal.' The well-known story of the latter being found, as a lad of twelve, tracing out on the hall floor geometrical propositions which he had evolved in his own head is not to be wondered at, nor yet that at sixteen he wrote a small work on Conic Sections, when one reflects upon the fame he was to win as a philosopher and writer, as well as a mathematician, in his too brief life of thirty-nine years. He was born in 1623 of a distinguished French family, and for the last half of his life he suffered from the effects of a serious disease which contributed to turn his attention from mathematics to religion and philosophy.

We learn from Todhunter how a certain gentleman of repute at the gaming tables set Pascal pondering on a question of probability concerning the fair division of stakes between two players who give up their game before its conclusion—an old problem cited in a work by Luca Pacioli as early as 1494. A correspondence followed between him and Fermat, then probably the two most distinguished mathematicians in Europe, and so began a science which has fascinated at one time or another all great mathematicians from that day to this.

The illustrious family of the Bernoullis, friends of Leibnitz, who championed his claim against that made by English mathematicians on behalf of Newton to the invention of the Calculus; De Moivre, an exile in England, owing to the revocation of the Edict of Nantes; Euler, Lagrange, and Laplace, who worked out in algebraical form Newton's theory of gravitation for the motion of the planets—all these had a share in building up the science of Probability, often by investigating problems in games of chance, where the conditions can be made mathematically perfect, so by careful analysis preparing the way for the use later of the same principles in matters of greater importance.

It has been said that the development of the subject owes more to Laplace (1749-1827) than to any other mathematician; nor did he confine himself to its theory: he would have earned fame by his astronomical applications alone. His method was to take certain observations, and to determine by means of probability whether the abnormalities present were merely the results of chance or whether there was some as yet undiscovered but constantly acting cause behind the phenomena observed. In this way he was led to highly interesting and important results such as those relating to the theory of the tides, the effect of the spheroidal shape of the earth on the motion of the moon, the irregularities of Jupiter and Saturn, and the laws which govern the motion of Jupiter's moons. It needs but a step in thought to pass from the discussion of such physical data to the statistics of social phenomena and the causes which determine abnormalities met with in that field. Professor Edgeworth, in making reference to books that have been written on *Probability* at the end of his excellent article under that heading in the *Encyclopædia Britannica*, remarks that 'as a comprehensive and masterly treatment of the subject as a whole, in its philosophical as well as mathematical character, there is nothing similar or second to Laplace's *Théorie analytique des probabilités*.']

CHAPTER XIII

SAMPLING (*continued*)—FORMULÆ FOR PROBABLE ERRORS

So far we have only considered the most simple case of random sampling when we take a sample of n independent events each of which falls into one of two classes according to its nature, the chance of entering either class being the same for every event: we have dealt, that is to say, more particularly with non-measurable characters. We pass on now to measurable characters which are distributed among several classes according to their size, so that a frequency distribution table can be set up for each sample; and assuming that the population from which the samples are drawn is homogeneous, the samples themselves containing each an adequate number of individuals, there should not be greater differences between one table and another than can be accounted for by random sampling. It is our object to discover how great such differences may be.

GENERAL POPULATION.

Class.	Frequency.
1st Group	Y_1
2nd Group	Y_2
..	..
kth Group	Y_k
..	..
..	..
	N

SAMPLE.

Class.	Frequency.
1st Group	y_1
2nd Group	y_2
..	..
kth Group	y_k
..	..
..	..
	n

Given a homogeneous population of N individuals which we will suppose could be distributed into a number of groups, Y_1 individuals in the first group, Y_2 in the second group, Y_3 in the third, and so on, according to the size of the organ or character under observation. Suppose a random sample of n individuals be taken from this population, and when they are assigned to their several groups let the frequency table now take the form shown, with y_1 individuals in the first group, y_2 in the second, and so on. *To find the probable error of y_k, the frequency observed in the kth group.*

Consider the selection of the n individuals, one by one in succession, to form the sample. When the first choice is made the probability that we shall get an individual falling into the kth group is, by definition, Y_k/N, and the probability will remain practically the same for each successive choice granted that N is considerable. We have thus n independent events, the chance of success (falling into the kth group) for each being $p(=Y_k/N)$ and the chance of failure being $q\left(=1-\frac{Y_k}{N}\right)$. The case is therefore analogous to the one previously considered to which the binomial distribution is applicable, so that the frequency to be expected in the kth group is np with S.D., $\sigma_{y_k}=\sqrt{npq}$; *i.e.* $y_k=np$ with a *p.e.*$=0{\cdot}6745\sqrt{npq}$.

Now in practice the numbers $Y_1, Y_2, Y_3 \ldots$ would not be known, and hence the true value of p would also be unknown, but since $y_k=np$, approximately, when the sample is of adequate size, we shall get a fair idea of the probable error involved by taking $p=y_k/n$, where y_k is the actual frequency observed in the kth group.

$$\text{Hence,}\quad \sigma^2_{y_k}=npq=y_k(1-p)=\mathbf{y}_k\left(1-\frac{\mathbf{y}_k}{\mathbf{n}}\right) \quad . \quad . \quad . \quad (1)$$

and the frequency in the kth group

$$=\mathbf{y}_k\pm0{\cdot}6745\sqrt{\mathbf{y}_k\left(1-\frac{\mathbf{y}_k}{\mathbf{n}}\right)} \quad . \quad . \quad . \quad (2)$$

The size of the S.D. is under ordinary conditions a test of the adequacy of the sample, for the frequency in the kth group, if due simply to random sampling, should not differ from its expected value by more than $3\sigma_{y_k}$ and σ_{y_k} should therefore be small compared with y_k itself.

Class.	Expected Frequency.	Observed Frequency.
1st Group	y_1	$y_1+\delta y_1$
2nd Group	y_2	$y_2+\delta y_2$
..	..	..
..	..	..
kth Group	y_k	$y_k+\delta y_k$
..	..	..
sth Group	y_s	$y_s+\delta y_s$
..	..	..
	n	n

To find the correlation between the frequencies in any two groups of a sample distribution.

Let the expected frequencies in the various groups of the sample be denoted by $y_1, y_2, \ldots, y_k, \ldots$, and suppose an error δy_k in y_k is associated with errors $\delta y_1, \delta y_2, \ldots, \delta y_s, \ldots$ in $y_1, y_2, \ldots, y_s, \ldots$ We require then the correlation between y_k and y_s.

Now although the group frequencies may change relative to one another, the total sum of frequencies in all groups is not affected, because the n individuals of the sample make up its composition in each case : to keep n constant the group frequencies must adjust themselves accordingly, which explains the correlation between them. Hence to compensate for an excess, δy_k (assuming $\delta y_k +^{ve}$), of frequency in any one group there must be a defect $(-\delta y_k)$ shared among the other groups, and the fairest way of sharing will be in proportion to the expected frequencies in the several groups.

But the total frequency divided between groups other than the kth is $(n-y_k)$, so that the proportion of $(-\delta y_k)$ due to the sth group is $y_s/(n-y_k)$, thus

$$\delta y_s = \frac{y_s}{n-y_k}(-\delta y_k).$$

Therefore,

$$\delta y_k \cdot \delta y_s = -y_s \cdot \delta y^2{}_k/(n-y_k)$$

$$= -\frac{y_s}{n} \cdot \frac{\delta y^2{}_k}{y_k\left(1-\frac{y_k}{n}\right)} \cdot y_k$$

$$= -\frac{y_s}{n} \cdot \frac{\delta y^2{}_k}{\sigma^2{}_{y_k}} \cdot y_k \quad . \quad . \quad . \quad (3)$$

by (1).

FIRST SAMPLE.

Size of Organ or Character observed.	Frequency of Observations.	First Moment.	Second Moment.
x_1	y_1	x_1y_1	$x^2{}_1y_1$
x_2	y_2	x_2y_2	$x^2{}_2y_2$
..	..	..	..
..	..	..	..
x_k	y_k	x_ky_k	$x^2{}_ky_k$
..	..	..	..
..	..	..	..
..	..	..	..
	n	$\Sigma(xy)$	$\Sigma(x^2y)$

This gives the product moment of the deviations from y_k and y_s in one particular sample ; summing for all such samples, remembering that by definition the coefficient of correlation between y_k

and y_s is $r_{v_k v_s} = \Sigma(\delta y_k \cdot \delta y_s)/\nu\sigma_{v_k}\sigma_{v_s}$, where ν is the total number of samples, also $\sigma^2_{v_k} = \Sigma\delta y^2_k/\nu$, we have

$$\nu r_{v_k v_s}\sigma_{v_k}\sigma_{v_s} = -\frac{y_s}{n} \cdot \nu \cdot y_k.$$

Therefore,
$$r_{y_k y_s} = -\frac{1}{n} \cdot \frac{y_k y_s}{\sigma_{y_k}\sigma_{y_s}} \quad . \quad . \quad . \quad . \quad (4)$$

gives the correlation required.

To find the p.e. of the mean of a sample of n observations. Let a frequency table be drawn up in the usual manner showing the number of observations y_1, y_2 . . . corresponding to organs of different sizes x_1, x_2 . . .

The mean referred to some fixed point as origin is then given by

$$M = (x_1 y_1 + x_2 y_2 + \ . \ . \ .)/n\ ;$$

also the mean square deviation of the sample referred to the same fixed point is μ^2_2, say, given by

$$\mu^2_2 = (x^2_1 y_1 + x^2_2 y_2 + \ . \ . \ .)/n,$$

and
$$\mu^2_2 - M^2 = \sigma^2$$

where σ is the S.D. of the sample.

For another sample of the same size the frequency distribution

SECOND SAMPLE.

Size of Organ or Character observed.	Frequency of Observations.	*First Moment.*
x_1	$y_1 + \delta y_1$	$x_1(y_1 + \delta y_1)$
x_2	$y_2 + \delta y_2$	$x_2(y_2 + \delta y_2)$
..	..	..
..	..	..
x_k	$y_k + \delta y_k$	$x_k(y_k + \delta y_k)$
..	..	..
..	..	..
..	..	..
	n	$\Sigma x(y + \delta y)$

may be slightly different, say, $y_1 + \delta y_1$, $y_2 + \delta y_2$, . . ., and consequently the mean will also be different, say,

$$M + \delta M = [x_1(y_1 + \delta y_1) + x_2(y_2 + \delta y_2) + \ . \ . \ . \ ]/n,$$

and, by subtraction,

$$\delta M=(x_1\delta y_1+x_2\delta y_2+\ .\ .\ .)/n \quad . \quad . \quad . \quad (5)$$

Now we want to determine the S.D. of the different values of M found among the different samples, and that is given by

$$\sigma^2_M=\Sigma\delta M^2/\nu,*$$

where Σ denotes summation for all samples and ν is the number of samples. This suggests that we should square both sides of equation (5), getting

$$n^2\ .\ \delta M^2=x^2_1\delta y^2_1+\ .\ .\ .\ +2x_1x_2\delta y_1\delta y_2+\ .\ .\ .$$

Therefore, $n^2\ .\ \nu\sigma^2_M=x^2_1\nu\sigma^2_{y1}+\ .\ .\ .\ +2x_1x_2\left(-\frac{y_1y_2}{n}\ .\ \nu\right)+\ .\ .\ .,$

by (3). Hence, making use also of (1),

$$n^2\sigma^2_M=x^2_1y_1\left(1-\frac{y_1}{n}\right)+\ .\ .\ .\ -\frac{2x_1y_1\ .\ x_2y_2}{n}-\ .\ .\ .$$

$$=(x^2_1y_1+\ .\ .\ .\ )-\frac{1}{n}(x^2_1y^2_1+\ .\ .\ .\ +2x_1y_1\cdot x_2y_2+\ .\ .\ .)$$

$$=n\mu^2_2-\frac{1}{n}(x_1y_1+\ .\ .\ .)^2$$

$$=n(\mu^2_2-M^2).$$

Thus $\quad \sigma^2_M=(\mu^2_2-M^2)/n=\sigma^2/\mathbf{n},$

and the probable error of the mean$=0{\cdot}6745\sigma/\sqrt{n}$ $\quad . \quad . \quad . \quad (6)$

The *p.e.* in the arithmetic mean found by taking a random sample of n events is a measure, so to speak, of the failure to hit the absolute mean, and it follows that the *precision* of the sample, the accuracy of aim at the mean, would be not unfairly measured by some quantity proportional to the reciprocal of the above expression, namely, $\sqrt{n}/0{\cdot}6745\sigma$. With such a measure the precision would evidently be increased if the number of observations in the sample were increased, being proportional to the square root of their number.

[It is desirable to draw a distinction here between what have been termed *biassed errors* and *unbiassed errors* ; errors due to random sampling are of the second class for there is, by hypothesis, no

[* We do not know the true mean for the population as a whole, but we take in place of it M, the value given by the sample, which we may do with little error if n is large. Similarly σ is the S.D. of the sample.]

reason why they should be in one direction rather than in another. Biassed errors, however, all tend to be in the same direction and they may arise in different ways, *e.g.* they may be due to faults of omission or commission on the part of the observer himself: he observes either carelessly or badly, omitting certain factors which ought to be taken into account, or so measuring or classifying his results that they appear always larger or less than they really are in fact.

Sometimes, although the bias is known to exist, it may be impossible to correct it: the most one can do is to bear it in mind and allow for it in using the results. A familiar example of this occurs in the collection of household budgets from the poor to find their standard of living, where it is only possible to get particulars from the more intelligent and thrifty class among them.

Whereas in the case of unbiassed errors due to random sampling we can diminish the probable error of the average by increasing the number of observations, the same is not true of errors which are biassed, for suppose an error ϵ in excess be made in each of n observations x_1, x_2, . . . x_n, the effect upon the average is to increase it from

$$\frac{x_1+x_2+\ .\ .\ .\ +x_n}{n} \text{ to } \frac{(x_1+\epsilon)+(x_2+\epsilon)+\ .\ .\ .\ +(x_n+\epsilon)}{n}.$$

i.e. from

$$\frac{x_1+x_2+\ .\ .\ .\ +x_n}{n} \text{ to } \frac{x_1+x_2+\ .\ .\ .\ +x_n}{n}+\epsilon,$$

so that the average is over-estimated by precisely the same amount. If, therefore, we know that bias exists, it is well, if possible, to correct it in each observation, for by so doing we change biassed into unbiassed errors, and though our corrections may be somewhat wide of the mark, the resultant error will then be diminished by increasing the number of observations: *e.g.* a farmer offers 400 sheep for sale and, being anxious to make a good bargain, he asks a higher figure for them than he is in reality prepared to take; let us suppose that this excess is 2s. 6d. for each sheep, then clearly the *average* price per sheep at which he is prepared to sell will be less than the amount he asks by 2s. 6d. also. But now suppose the buyer, a simple person knowing little of the prices of sheep and less of the ways of men, goes through the flock one by one and makes the error of offering a price either much above or much below what the seller is prepared to take; even if his unbiassed offers

differ by as much as 10s. for each sheep from the seller's reserve price, so long as they are random in direction, *i.e.* sometimes too much and sometimes too little, the resultant difference in the *average* from what the seller is prepared to take will probably not greatly exceed ⅔10s./$\sqrt{400}$, or 4d. per sheep.

We can sometimes diminish the effect of bias, even when its extent is unknown, by working with the ratios of the quantities affected instead of with the quantities themselves: *e.g.* suppose biassed errors, ϵ_1 and ϵ_2, enter into the measurement of the variables x_1 and x_2, both in excess, the ratio of the variables then

$$
\begin{aligned}
&=(x_1+\epsilon_1)/(x_2+\epsilon_2)\\
&=x_1\left(1+\frac{\epsilon_1}{x_1}\right)\Big/x_2\left(1+\frac{\epsilon_2}{x_2}\right)\\
&=\frac{x_1}{x_2}\left(1+\frac{\epsilon_1}{x_1}\right)\left(1+\frac{\epsilon_2}{x_2}\right)^{-1}\\
&=\frac{x_1}{x_2}\left(1+\frac{\epsilon_1}{x_1}\right)\left(1-\frac{\epsilon_2}{x_2}+\text{higher powers of }\epsilon_2\right)\\
&=\frac{x_1}{x_2}\left(1+\frac{\epsilon_1}{x_1}-\frac{\epsilon_2}{x_2}\right),
\end{aligned}
$$

if we omit higher powers of ϵ_1 and ϵ_2 than the first on the understanding that they are both comparatively small. Suppose, for example, there was an error of 5 per cent. made in measuring x_1 and an error of 3 per cent. of like sign in measuring x_2 then the resulting error in x_1/x_2 would be 5 per cent.$-$3 per cent.$=$2 per cent. Clearly the same holds good also if the errors are both in defect. This explains why a comparison of results arranged, say, on the index number principle may be trustworthy, although the method of formation of the numbers themselves may be in some respects faulty, granted that the same faults are repeated each year so as to produce like errors, *i.e.* the bias is to be unchanged in character. To correct the faults in one case and not in the other would prejudice the success of the method, since it depends upon the errors counteracting one another.]

Example (1).—To illustrate the important result we have obtained for the p.e. of the mean of n observations let us return to the experiment of selecting 900 random digits. The distribution actually obtained, and the theoretical distribution to be expected in the

long run if the experiment were repeated several hundred times and the average taken, are shown in the following table :—

TABLE (37). DISTRIBUTION OF 900 RANDOM DIGITS.

Digit.	Frequency Observed.	Theoretical Frequency.	Digit.	Frequency Observed.	Theoretical Frequency.
0	95	90	5	80	90
1	96	90	6	82	90
2	93	90	7	72	90
3	105	90	8	90	90
4	91	90	9	96	90

It is a simple matter to calculate the mean and S.D. for the distribution from this table in the usual way ; the results are :—

Observed mean$=4{\cdot}38$; S.D.$=2{\cdot}911$
Theoretical mean$=4{\cdot}50$; S.D.$=2{\cdot}872$.

Thus the p.e. of the mean based on the sample

$$=\pm 0{\cdot}6745\times 2{\cdot}911/\sqrt{900}$$
$$=\pm 0{\cdot}065,$$

and 4·38 differs from 4·50 by less than three times the p.e.

The 36 averages of samples of 25 events apiece were also calculated, and the following were the results obtained :—

2·76, 3·32, 3·68, 3·72, 3·72, 3·72, 3·76, 3·80, 3·92, 3·92, 4·08, 4·12, 4·16, 4·16, 4·16, 4·28, 4·36, 4·40, 4·40, 4·40, 4·44, 4·60, 4·64, 4·68, 4·72, 4·72, 4·76, 4·88, 4·96, 5·00, 5·00, 5·00, 5·08, 5·28, 5·40, 5·72.

The mean of this distribution$=157{\cdot}72/36=4{\cdot}381$, and the S.D.$=0{\cdot}612$. But the S.D. of the whole distribution of 900 digits $=2{\cdot}911$, and therefore the S.D. of the distribution of averages of samples of 25 digits should be $2{\cdot}911/\sqrt{25}=0{\cdot}582$, differing from 0·612 by about 5 per cent.

To find the p.e. of the sum or difference of two variables. Let the mean values of the two variables be denoted by y and z, so that deviations from these values found in a particular sample may be denoted by δy and δz. If then we write

$$u=y+z$$

we have

$$\delta u=\delta y+\delta z. \quad . \quad . \quad . \quad (7)$$

To find the S.D. of u we therefore require $\Sigma(\delta u^2)/\nu$, where the Σ denotes summation for all samples and ν is the number of samples.

But, squaring both sides of equation (7), we have

$$\delta u^2 = \delta y^2 + \delta z^2 + 2\delta y \delta z.$$

Thus $$\Sigma\delta u^2 = \Sigma\delta y^2 + \Sigma\delta z^2 + 2\Sigma(\delta y \delta z),$$

where the summation extends to all samples. Hence

$$\nu\sigma^2_u = \nu\sigma^2_y + \nu\sigma^2_z + 2\nu\sigma_y\sigma_z r_{yz}$$

or $$\sigma^2_u = \sigma^2_y + \sigma^2_z + 2r_{yz}\sigma_y\sigma_z$$

where r_{yz} is the correlation between the variables. And the p.e. $= 0{\cdot}6745\sigma_u$.

The p.e. of the difference of two variables follows at once by changing the sign of z throughout; for, if

$$v = y - z,$$

we have $$\delta v^2 = \delta y^2 + \delta z^2 - 2\delta y \delta z,$$

and $$\sigma^2_v = \sigma^2_y + \sigma^2_z - 2r_{yz}\sigma_y\sigma_z.$$

Generally, if x_1, x_2, . . . x_n be the mean values of n variables, and if δx_1, δx_2, . . . δx_n denote deviations from these values in a particular sample, we may write

$$u = x_1 + x_2 + \ . \ . \ . \ + x_n$$

and $$\delta u = \delta x_1 + \delta x_2 + \ . \ . \ . \ + \delta x_n.$$

Thus $$\Sigma\delta u^2 = \Sigma\delta x_1{}^2 + \ . \ . \ . \ + 2\Sigma(\delta x_1 \delta x_2) + \ . \ . \ .$$

whence $$\sigma^2_u = \sigma^2_{x_1} + \ . \ . \ . \ + 2r_{x_1x_2}\sigma_{x_1}\sigma_{x_2} + \ . \ . \ .$$

Important Corollary. If y and z are quite independent so that r_{yz} is zero, the p.e. of their sum and the p.e. of their difference have the same value, namely, the square root of the sum of the squares of the p.e.'s of y and z themselves, which

$$= 0{\cdot}6745\sqrt{(\sigma^2_y + \sigma^2_z)} \quad . \quad . \quad . \quad (8)$$

This result is exceedingly important, because it can be directly used to test whether a difference between two samples is accidental, *i.e.* whether it is such as might arise through sampling, or whether it implies a real difference between the two populations from which the samples are selected. An example will illustrate the procedure :—

Example (2). In a study of *Minimum Rates in the Tailoring Industry,* by R. H. Tawney, a table is given (p. 114) which suggests

that 'in the north of England women work in the tailoring trade when they are young . . . in London and Colchester they have to work when they are older.' Taking some figures from that table we find :—

District.	Workers over 35 years old.	Workers at all ages.	Proportion over 35.
London and Essex .	11,718	35,316	0·332
Manchester and Leeds .	4,029	21,822	0·185

The difference between the proportions over 35 years of age

$$=(0{\cdot}332-0{\cdot}185)=0{\cdot}147.$$

Let us suppose for the moment that this difference is not significant of any real difference in conditions between the two districts, but is merely due to random sampling. In that case the most natural value to assign to the true proportion of women workers over 35 for the trade as a whole, as given by these figures, would be

$$p=\frac{11{,}718+4{,}029}{35{,}316+21{,}822}=\frac{15{,}747}{57{,}138}=0{\cdot}276.$$

The S.D. for the first sample (London and Essex) would then be

$$\sigma_1=\sqrt{(pq/n)}=\sqrt{[0{\cdot}276\times0{\cdot}724/35{,}316]},$$

and for the second sample (Manchester and Leeds) would be

$$\sigma_2=\sqrt{[0{\cdot}276\times0{\cdot}724/21{,}822]}.$$

Hence the p.e. for the difference between the proportions in the two samples would be roughly

$$=\tfrac{2}{3}\sqrt{(\sigma^2_1+\sigma^2_2)},\text{ by (8)},$$
$$=\tfrac{2}{3}\sqrt{[0{\cdot}276\times0{\cdot}724(1/35{,}316+1/21{,}822)]}$$
$$=\tfrac{2}{3}\sqrt{[0{\cdot}276\times0{\cdot}724/13500]}$$
$$=0{\cdot}0026.$$

The actual difference between the proportions, 0·147, being much more than 3(0·0026), is certainly significant of a greater difference between the two populations than can be explained by random sampling alone.

Another method of attack would be to assume a real difference between the two populations, if other considerations led us to suspect such a difference, and to find whether such a difference could be disguised by random sampling. In that case the proper proportion to assume for the first sample would be 0·332, giving

$$\sigma_1=\sqrt{[0{\cdot}332\times0{\cdot}668/35{,}316]}=\sqrt{628/10^4},$$

and for the second sample the proportion would be 0·185, giving

$$\sigma_2=\sqrt{[0{\cdot}185\times0{\cdot}815/21{,}822]}=\sqrt{691/10^4}.$$

Hence the p.e. for the difference between these two proportions due to random sampling would be

$$=\tfrac{2}{3}\sqrt{(\sigma_1^2+\sigma_2^2)},\text{ by (8)},$$
$$=\tfrac{2}{3}\frac{1}{10^4}\sqrt{(628+691)}$$
$$=0{\cdot}0024.$$

The actual difference is 0·147, which certainly could not be outbalanced by an error in the opposite direction due to random sampling, because it is much more than three times the probable error due to sampling.

Sometimes we have to test the difference, not between two simple proportions, but between two sample distributions. In that case the mean of each sample may be calculated so that the difference (M_1-M_2) between the means is known; to find out whether or not it is significant of some real difference between the two populations from which the samples are drawn, (M_1-M_2) is compared with its p.e., namely

$$0{\cdot}6745\sqrt{(\sigma^2_{M1}+\sigma^2_{M2})},$$

or

$$0{\cdot}6745\sqrt{(\sigma^2_1/n_1+\sigma^2_2/n_2)} \quad . \quad . \quad . \quad (9)$$

where n_1 and n_2 are the numbers of observations in the two samples respectively, and σ_1, σ_2 are the S.D.'s of the samples. Unless (M_1-M_2) is definitely greater than some two or three times this expression we cannot be very sure that the difference between M_1 and M_2 may not have arisen merely through random sampling, and it may quite likely not be significant * of any real difference between the two populations as regards the organ or character which is under consideration.

[* It should be observed that the S.D. provides a wider margin for significance than the p.e., because a range of approximately 3 p.e. $=3\cdot\tfrac{2}{3}\sigma=2\sigma$ only. It is quite safe therefore to attach no great significance to a difference which does not exceed two or three times the p.e.]

Example (3).—Statistics have been collected to test whether there is any significant difference between the eggs laid in general by cuckoos and those laid by them in the nests of particular species of foster parents. Results of the following kind were obtained [see *Biometrika*, vol. iv., pp. 363-373, *The Egg of Cuckulus Canorus* (2nd Memoir), by O. H. Latter]:—

Group.	Number of Eggs.	Mean Length (mms.)	S.D. (mms.)	Significance Test.	Remarks.
Eggs of the Cuckoo race in general	1572	22·3	0·9642	..	..
Eggs laid in nests of					
Garden Warbler	91	21·9	0·7860	7·0	Significant.
White Wagtail	115	22·4	0·7606	1·6	Not significant.
Hedge Sparrow	58	22·6	0·8759	3·75	Probably significant.

The difference between the mean lengths of eggs laid in the nests of garden warblers and those laid by cuckoos in general

$$=22{\cdot}3-21{\cdot}9=0{\cdot}4 \text{ mms.}$$

The p.e. of this difference

$$=0{\cdot}6745\sqrt{[(0{\cdot}7860)^2/91+(0{\cdot}9642)^2/1572]}, \text{ by (9)},$$
$$=0{\cdot}6745\sqrt{(0{\cdot}007380)}$$
$$=0{\cdot}058.$$

Hence the significance test

$$=0{\cdot}4/0{\cdot}058=7{\cdot}0,$$

and we conclude that the difference in length between the two classes of eggs is certainly significant. Similarly the other cases may be tested.

In the example just given, to find out whether one population differed from another, the arithmetic means have been compared; but the mean alone will scarcely serve to establish the identity of any population. For example, we can conceive of two distinct races of men, both of the same mean height, but one race embracing a number of giants and dwarfs. Of course if we agreed to *define* two races as identical when they have the same mean heights, there would be nothing more to be said, but that would certainly only be a very rough-and-ready attempt at classification.

Taking into consideration only the character of height, a further step in definition would be to measure the mode or most fashionable

height, and the dispersion or variability—absolute: the standard deviation, and relative: the coefficient of variation—of the two races. Then, after comparing heights with sufficient detail, the attention could be turned to innumerable other characters, skull and body measurements, physical, mental and even moral attributes.

Clearly the difficulty of definition and of establishment of identity grows as we pass along the scale from physical to moral. Moreover, other statistical constants must be requisitioned when the question of the existence and degree of relationship between two organs or characters is to be determined. As the S.D. and the C. of V. serve to measure the amount of variability, so the coefficient of correlation comes in to measure the amount of likeness or association. Further, and especially in problems of inheritance, the coefficient of regression must be measured. It might seem at first sight hopeless to try and measure the correlation between two such characters as athletic capacity and health in the same boy, or between the truthfulness of one boy and that of his brother; but the genius of Karl Pearson has gone some way to solve even this difficult problem by means of a system of adjectival instead of numerical classification [see *Phil. Trans.*, vol. 195A, pp. 1-47, *On the Correlation of Characters not Quantitatively Measurable*, and, as an exceptionally interesting application of the method, see Pearson, *On the Laws of Inheritance in Man*, ii.; *On the Inheritance of the Mental and Moral Characters in Man and its Comparison with the Inheritance of the Physical Characters*; *Biometrika*, vol. iii. pp. 131-190]. In short, for a full and exact definition of a population of any kind, human or otherwise, it is necessary to measure not only the means, but all the more important statistical constants, modes, medians, S.D.'s, C.'s of V., coefficients of correlation and regression, and so on, and it is no less necessary to calculate also their probable errors if we are to test the real significance of such differences as are observed in these constants between two samples from the same or from different populations.

The probable errors for the more important constants, some of which are only introduced later in the book, are collected together in Table (38) for reference. The proofs in general are a little intricate and would be lacking in interest to the ordinary person, who is satisfied to take algebraical analysis on trust so long as he understands the nature of the results he uses, but the more mathematical reader who is anxious to see proofs may refer for some of them to *Biometrika*, vol. ii., pp. 273-281, *Editorial*, *On the Probable Errors*

of Frequency Constants, which has been freely consulted on the subject here.

The usual notation is adopted, n being the total number of observations in the given distribution, supposed normal in general, σ the S.D., etc.

TABLE (38). PROBABLE ERRORS OF STATISTICAL CONSTANTS.

Statistical Constant.	Probable Error (=0·6745 S.D.).
Any observed group frequency, y	$0{\cdot}6745 \times \sqrt{[y(1-y/n)]}$
The mean of a distribution of any type	„ $\sigma/\sqrt{n}$
The S.D. of a normal distribution, σ	„ $\sigma/\sqrt{2n}$
The second moment about the mean, μ_2	„ $\sigma^2\sqrt{2/n}$
„ third „ „ „ μ_3	„ $\sigma^3\sqrt{6/n}$
„ fourth „ „ „ μ_4	„ $\sigma^4\sqrt{96/n}$
The coefficient of variation, v	„ $\frac{v}{\sqrt{2n}}\left[1+2\left(\frac{v}{100}\right)^2\right]^{\frac{1}{2}}$
The coefficient of correlation, r	„ $(1-r^2)/\sqrt{n}$
The correlation ratio, η	„ $(1-\eta^2)/\sqrt{n}$, nearly
X, as determined from $(X-\bar{X})=r\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$, when Y is given	„ $\sigma_x\sqrt{(1-r^2)}$
Y, as determined from $(Y-\bar{Y})=r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$, when X is given	„ $\sigma_y\sqrt{(1-r^2)}$
Distance between mode and mean in a skew distribution	„ $\sigma\sqrt{(3/2n)}$
Skewness	„ $\sqrt{(3/2n)}$
β_2 (which should = 3 for a normal distribution)	„ $\sqrt{(24/n)}$
β_1 („ „ = 0 „ „)	„ 0
$\sqrt{\beta_1}$	„ $\sqrt{(6/n)}$

Example (4).—In the example which follows are given data necessary for testing the significance of differences in variability as well as in mean values. They represent an attempt made to find whether members of a particular species of crab caught in shallow water differed with regard to certain characteristics from those caught in comparatively deep water [see *Biometrika*, vol. ii., pp. 191 *et seq.*, *Variation in Eupagurus Prideauxi*, by E. H. J. Schuster]. Only a few of the results are recorded here, to two decimal places ; the reader will find it a valuable exercise to verify for himself the p.e.'s given in each case.

Measurement Made.	Sex.	*Locality.*	Mean (mm.).	S.D. (mm.).	C. of V. per cent.
Carapace length	Male	Deep water	8·59±0·05	1·67±0·04	19·45±0·44
,, ,,	,,	Shallow ,,	8·41±0·04	1·49±0·03	17·75±0·37
,, ,,	Female	Deep ,,	7·54±0·03	0·94±0·02	12·49±0·28
,, ,,	,,	Shallow ,,	7·12±0·02	0·86±0·02	12·12±0·25

Difference of Means (mm.).	Difference of S.D.'s (mm.).	Difference of C.'s of V. per cent.	Sex.
0·18±0·07 (poss. sig.)	0·18±0·05 (prob. sig.)	1·70±0·58 (poss. sig.)	Male
0·42±0·04 (sig.)	0·08±0·03 (poss. sig.)	0·37±0·37 (not sig.)	Female

The significance or otherwise of differences between variabilities in the case of cuckoos' eggs (p. 161) might be tested in the same way.

CHAPTER XIV

FURTHER APPLICATIONS OF SAMPLING FORMULÆ

WE have been discussing in the last chapter how to test two samples, supposed each to contain homogeneous material, to find out whether they belong to the same or to different types of population, but the further question often arises as to whether a sample is or is not homogeneous.

Example (1).—To this we may obtain a partial answer by working out the statistical constants of the sample and their p.e.'s in order to compare them with the corresponding constants for a sample or series of samples believed to be homogeneous and of the same type. For example, Professor Karl Pearson has measured the skulls of skeletons of the Naqada race, excavated in Upper Egypt by Professor Flinders Petrie and presumed to be some 8000 years old, and he places his results for comparison alongside those for certain other races admittedly homogeneous [see *Biometrika*. vol. ii., p. 345, *Homogeneity and Heterogeneity in Collections of Crania*] :—

Series.		Number of Observations.	Variability (mm.).	
			Skull Length.	Skull Breadth.
Skulls	Ainos . . .	76	5·936	3·897
	Bavarians . . .	100	6·088	5·849
	Parisians . . .	77	5·942	5·214
	Naqadas . . .	139	5·722	4·612
	English . . .	136	6·085	4·976
Living heads	Cambridge undergrad'tes	1000	6·161	5·055
	English criminals .	3000	6·046	5·014
	Oraons of Chota Nagpur	100	5·916	4·397
Mean Variability . . .			5·987	4·877

The S.D. of the variability of skull length calculated from this series $=0\cdot129$ mm. and of the variability of skull breadth $=0\cdot545$ mm., and these supply standards for valuing the differences between the Naqada and the mean variabilities.

Another method of procedure is to take a random sample out of the sample itself, assuming the latter is large enough to admit of an adequate sub-sample, and to compare the constants of the whole and part. When they do not differ beyond the limits allowed by random sampling the inference is that the whole may be treated as a homogeneous class if judged by this test alone.

Example (2).—In an interesting and important memoir, *On Criminal Anthropometry and the Identification of Criminals*, by W. R. Macdonell [*Biometrika*, vol. i., pp. 177 *et seq.*], the author uses this method to test the homogeneity of a class of 3000 criminals by measuring also a random sample of 1306 criminals out of the 3000. He obtained, for example,

S.D. of head length $=6\cdot04593\pm0\cdot05265$ mm., for the 3000 criminals ;
" " $=6\cdot00247\pm0\cdot07922$ " " 1306 "

The difference between the variabilities in the sample and sub-sample, by result (8) on p. 158,

$$=0\cdot04346\pm\sqrt{[(0\cdot05265)^2+(0\cdot07922)^2]}$$
$$=0\cdot04346\pm0\cdot09512$$

which is certainly not significant. If the same holds good with regard to the means and other constants, then the whole may be said to be homogeneous so far as this test goes.

Example (3).—Another example may be given from the memoir on *Variation and Correlation in Brain Weight*, by Raymond Pearl, [*Biometrika*, vol. iv., pp. 13 *et seq.*]. The author wished particularly to investigate the change of brain weight with age ; on the hypothesis that the weight of the brain reaches a maximum between the ages of 15 and 20, remains unchanged from 20 to 50, and then begins to decline and so continues till death, the material was divided into a 'Young' series, ages 20 to 50, and a 'Total' series including all between 20 and 80. The 'Young' series thus formed a selection from the 'Total' series, but a selection based on age and not on brain weight. If there were no correlation between age and brain weight, this selection, based as it is on age, would, of course, be random as regards brain weight. Now correlation does exist between the two, but it is so slight that, within the limits

of error, the 'Young' series does form practically a random sample of the 'Total' series, as is shown by the following figures :—

DIFFERENCE IN VARIATION CONSTANTS BETWEEN YOUNG AND TOTAL SERIES (WRITTEN WITH A POSITIVE SIGN WHEN THE YOUNG SERIES GIVES THE GREATER VALUE).

	Male.		Female.	
	S.D.	C. of V.	S.D.	C. of V.
Swedes	+2·851±4·066	+0·122±0·291	+ 4·786±5·465	+0·271±0·435
Bavarians	−1·888±3·556	−0·173±0·234	−10·357±3·909	−0·941±0·320

Thus in only one case, that of the Bavarian females, is the difference between the variabilities, S.D. or C. of V., of the two series as great as its probable error, and even in that case the differences, 10·357 and 0·941, are not three times as large as their respective p.e.'s, 3·909 and 0·320. Dr. Pearl concludes from these and similar results that 'the series are reasonably homogeneous in other respects than age.'

The reader is recommended to test his knowledge of the formulæ for probable errors by applying them to the following examples. Dr. Alice Lee, in a note on *Dr. Ludwig on Variation and Correlation in Plants* [*Biometrika*, vol. i., p. 316] makes use of the statistics relating to Ficaria Verna in Example (4). Those in Example (5) are taken from among a large number of others in the highly interesting memoir, *On the Laws of Inheritance in Man*, by Professor Karl Pearson and Dr. Alice Lee [*Biometrika*, vol. ii., pp. 357 *et seq.*] cited once before.

Example (4).—VARIATION AND CORRELATION IN FICARIA VERNA.

No. of Observations.	Mean No. of Petals ; S.D.	Mean No. of Sepals ; S.D.	Correlation between No. of Sepals and No. of Petals.
1000 (Greiz A) .	8·286 ; 1·3382	3·695 ; 0·8524	0·2439±0·0201
1000 (Greiz G) .	8·232 ; 0·9954	3·437 ; 0·7033	0·2480±0·0200

We have here all the data necessary to find the p.e.'s of the means, variabilities, and correlations, and we wish to know whether

the differences between the means and variabilities of the A and G plants can be accounted for by random sampling alone.

For example, the difference between the petal means

$$=(8\cdot286-8\cdot232)\pm\tfrac{2}{3}\sqrt{\left[\frac{(1\cdot3382)^2}{1000}+\frac{(0\cdot9954)^2}{1000}\right]}$$

$$=0\cdot054\pm0\cdot035.$$

Clearly this difference, being not so great as twice its p.e., is not significant and may quite well be due to random sampling.

Again, the difference between the petal variabilities

$$=(1\cdot3382-0\cdot9954)\pm\tfrac{2}{3}\sqrt{\left[\frac{(1\cdot3382)^2}{2000}+\frac{(0\cdot9954)^2}{2000}\right]}$$

$$=0\cdot3428\pm0\cdot025$$

which is certainly much too great to be explained away by random sampling merely.

Similarly the differences between the sepal means, between the sepal variabilities, and between the correlations, may be tested for significance by comparison with their p.e.'s.

Example (5).—Size and Variability of Stature in the Two Generations.

	Father.	Mother.	Son.	Daughter.
Mean height (in.)	$67\cdot68\pm0\cdot06$	$62\cdot48\pm0\cdot05$	$68\cdot65\pm0\cdot05$	$63\cdot87\pm0\cdot05$
S.D. (in.) . .	$2\cdot70\pm0\cdot04$	$2\cdot39\pm0\cdot04$	$2\cdot71\pm0\cdot04$	$2\cdot61\pm0\cdot03$
C. of V. (per cent.)	$3\cdot99\pm0\cdot06$	$3\cdot83\pm0\cdot06$	$3\cdot95\pm0\cdot06$	$4\cdot09\pm0\cdot05$

The student in this case might use one of the formulæ for the p.e.'s to find the number of fathers, mothers, sons, or daughters observed when the p.e.'s are known, and then the remaining p.e.'s might be verified when the numbers of observations are found.

As evidence of 'assortative mating,' the tendency of like to mate with like, the following particulars are given, based on 1000 to 1050 cases of husband and wife :—

Correlation between	stature	of husband	and	stature	of wife	$=0\cdot2804\pm0\cdot0189$
,, ,,	span	,, ,,	,,	span	,, ,,	$=0\cdot1989\pm0\cdot0204$
,, ,,	forearm	,, ,,	,,	forearm	,, ,,	$=0\cdot1977\pm0\cdot0205$

To measure the average intensity of inheritance, the extent of

resemblance between parents and children in any character, coefficients of correlation are calculated such as the following :—

Coefficient of Correlation	
between stature of father and stature of son	$=0{\cdot}514\pm0{\cdot}015$
,, ,, ,, ,, ,, ,, ,, daughter	$=0{\cdot}510\pm0{\cdot}016$
,, ,, ,, mother ,, ,, ,, son	$=0{\cdot}494\pm0{\cdot}016$
,, ,, ,, ,, ,, ,, ,, daughter	$=0{\cdot}507\pm0{\cdot}016$

[In verifying the p.e.'s for this case take the number of observations to be 1024.]

One more extract may be quoted, a prediction table, giving the probable mean stature of sons of fathers of given stature, and so on :—

Son's probable stature	$=33{\cdot}73+0{\cdot}516$ (father's stature) $\pm1{\cdot}56$
Daughter's ,, ,,	$=30{\cdot}50+0{\cdot}493$ (,, ,,) $\pm1{\cdot}51$
Son's ,, ,,	$=33{\cdot}65+0{\cdot}560$ (mother's stature) $\pm1{\cdot}59$
Daughter's ,, ,,	$=29{\cdot}28+0{\cdot}554$ (,, ,,) $\pm1{\cdot}52$.

All values given in this example for the p.e.'s should be verified.

Before we consider further applications of these principles to questions of a somewhat different kind, let us imagine a very simple though artificial illustration. Suppose we have 999 sheep, each one ticketed, the numbers on the tickets running from 1 to 999. Also suppose 666 of these sheep are white and 333 are black, so that, if we pick out any one at random, the chance of it being black is 333/999 or 1/3. Let us call picking a black sheep a 'success,' then $p=1/3$, $q=2/3$.

We proceed now to select 99 sheep in succession at random from the flock with the understanding that each sheep is returned into the flock before the next is picked out. This insures that the chance of a success at each selection remains equal to 1/3 and, of course, there is nothing to prevent the same sheep being picked more than once. The selection might practically be made by placing in a box 999 tickets, numbered from 1 to 999, one to correspond to each sheep, then picking out 99 of them in succession, being careful to replace each and to shake up the box before picking out the next ; if there were absolutely no difference between the tickets, such as would cause one to be picked more easily than another, the selection made in this way would be random in the

sense required, and the tickets so chosen would determine which sheep were to be taken and which left.

The proportion of black sheep to be expected in such a random selection of 99 is 1/3, but, if we only perform the experiment once, it is quite likely that the proportion we actually get will differ from 1/3 by an amount

$$=0{\cdot}6745\sqrt{(pq/n)}$$
$$=0{\cdot}6745\sqrt{(\tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{1}{99})}$$
$$=1/31, \text{ about,}$$

while it is unlikely that the proportion will differ from 1/3 by much more than 3/31, or 1/10.

Conversely—and it is really the converse which is useful in practice—if we do not know the proportion of black sheep in the whole flock, we may get a fair estimate of it by taking a random sample of 99 sheep (any other number will serve the purpose, but the larger the better for accuracy), and if we find that in this sample there are 33 black sheep, *i.e.* $p=33/99=1/3$, it will appear that the value of p for the whole flock is 1/3, subject to a probable error $0{\cdot}6745\sqrt{(pq/n)}$ in excess or defect, *i.e.* the true proportion for the whole flock may quite likely differ from 1/3 by as much as 1/31, but it is unlikely to differ by much more than 1/10. It should be noticed that the calculation of the probable error in this converse case is based upon the value of p given by the sample taken, for that is the only value of which we have knowledge.

Too much stress can scarcely be laid on the fact that the samples chosen must be absolutely unbiassed, otherwise the use of the formulæ np and $\sqrt{(npq)}$, or the corresponding proportional formulæ, cannot be justified : each sheep in our illustration must have the same chance of being picked, and no one selection is to have any influence on another. The failure to appreciate this essential point has led to no little waste of time and effort in the collection of valueless statistics.

The method of sampling has been employed in a way at once interesting and useful by Dr. A. L. Bowley, and, as some of this work has barely received the attention it deserves, it may be well to explain two of his experiments in some detail.

The first was of interest because its results could be tested by an examination of the original record from which the sample was taken. The details concerning it are abstracted from the *Journal of the Royal Statistical Society*, September 1906.

Example (6).—Bowley sampled the dividends paid by 3878

companies as quoted in the *Investors' Record.* His sample consisted of 400 of these companies, *i.e.* about 10 per cent., selected in a purely arbitrary fashion thus: the investigator took a Nautical Almanac and noted down the last digits of one of the tables, recording them in groups of four, but if any particular group gave a number bigger than 3878 he rejected it. In this way each of the numbers between 1 and 3878 had an equal chance of selection (for numbers under four figures would appear like 0327, 0042, 0009, which would be taken to represent 327, 42, 9 respectively), and the selection of one had no influence on that of any other. The companies in the *Investors' Record* were numbered consecutively, and the dividends corresponding to the 400 arbitrary numbers obtained formed the sample with which Bowley worked.

After making some interesting deductions with regard to the average for the whole distribution, to which we shall return presently, he proceeded to forecast the grouping of the original companies as to their dividends by setting out the grouping discovered in the sample 400, as follows, using the standard deviation in place of the probable error as the error due to random sampling:—

TABLE (39). DISTRIBUTION OF DIVIDENDS PAID BY A SAMPLE OF 400 COMPANIES.

(1) Dividend.	(2) Sample of 400 Companies.	(3) Percentage of Sample Companies in each Class.	(4) Percentage of all Companies in each Class.
Nil.	28	7 with S.D. = 1·27	6
£1 to £2, 19s. 9d.	6	$1\frac{1}{2}$	1·5
£3 to £3, 9s. 9d.	37	$9\frac{1}{4}$,, = 1·46	8·4
£3, 10s. to £3, 19s. 9d.	71	$17\frac{3}{4}$,, = 1·90	18·8
£4 to £4, 9s. 9d.	64	16 ,, = 1·83	17·3
£4, 10s. to £4, 19s, 9d.	53	$13\frac{1}{4}$,, = 1·68	13·8
£5 to £5, 19s. 9d.	60	15 ,, = 1·78	17·7
£6 to £7, 19s. 9d.	48	12 ,, = 1·63	10·8
£8 to £10, 19s. 9d.	29	$7\frac{1}{4}$,, = 1·29	3·8
Above £11 . .	4	1	1·9

In col. (3) the S.D. for each group was calculated as follows:—for the first group: out of 400 possible events we have 28 successful events, meaning by 'successful' here 'a company paying no dividend,' thus

$$p = 28/400, \qquad q = 372/400.$$

Hence the S.D. of the frequency in the first group

$$=\sqrt{[28(1-\tfrac{28}{400})]}$$
$$=\sqrt{(28\times 372)}/20$$
$$=5{\cdot}1.$$

Since this is for a sample of 400, the S.D. of the *percentage** frequency in the first group

$$=\tfrac{1}{4}(5{\cdot}1)=1{\cdot}27.$$

The other S.D.'s are calculated in the same way, but when the number in a class is very small the forecast can scarcely be relied upon and consequently the S.D. is not inserted.

It will be noted, by comparing with the numbers in col. (4), showing the corresponding percentages for all the 3878 companies, that every forecast was remarkably good except one, class £8 to £10, 19s. 9d., where the error approaches three times the S.D., and the exception will serve as a warning that, in working with samples, the unexpected sometimes happens. Professor Edgeworth, in his Presidential Address to the Royal Statistical Society (1912), points out that the method appears to be a permanent institution in the Statistical Bureau at Christiania, where it has given very good results. These can be checked or ' controlled ' for safety if complete statistics are obtainable under some heads. He fairly sums up the utility of sampling when he says that ' we may obtain from samples a general outline of the facts—often sufficient for the initiation of a project like that of insurance—rather than the features in detail.'

Bowley also divided up his 400 random samples into 40 groups of 10 companies each, and calculated the average for each group. The S.D. for these 40 averages was found in the usual way, giving 0·775. But since this was the S.D. for averages of 10, we conclude that

$$(\text{the S.D. for the distribution of the 400 companies})/\sqrt{10}=0{\cdot}775$$

i.e. the S.D. for the distribution of the 400 companies $=0{\cdot}775\sqrt{10}$.

Hence, applying the same principle again,

the S.D. of the average of the 400 sample companies

$$=0{\cdot}775\sqrt{10}/\sqrt{400}$$
$$=£0{\cdot}122.$$

[* It would not be correct to take $\sqrt{[7(1-\tfrac{7}{100})]}$ as the S.D. of the percentage frequency in the first group; this value would be double the true value, namely, $\tfrac{1}{4}\sqrt{[28(1-\tfrac{28}{400})]}=\tfrac{1}{2}\sqrt{[7(1-\tfrac{7}{100})]}$, because the accuracy is increased by increasing the number of events in a sample, and the sample here is really 400 and not 100.]

Now the average of the 400 samples turned out to be £4·7435. Hence it was judged that, if this was a fair selection (and the random method adopted was such as to make it fair in all reasonable likelihood), the average for the 3878 companies should certainly lie between

$$£[4{\cdot}7435 \pm 3(0{\cdot}122)].$$

The true average was found by actual calculation to be £4·779, well within the above limits, although the original items varied from nil to £103, being grouped according to the nature of the security —Government, Railways, Mines, etc., etc., and the averages and S.D.'s on successive pages differed materially. This aggregation, Bowley remarks, is very similar to that found in wages in different occupations and localities, and in many other practical examples.

The value of the second experiment due to Dr. Bowley lies in the suggestion that similar means can be applied with good results to the investigation of many social phenomena.

If out of a large group a comparatively small sample of statistics is collected in the purely random manner already described, we are able by such means to estimate what is the average, and even to obtain limits between which the average will almost certainly lie, in the large group based upon values found for the average and S.D. in the small sample.

Example (7).—With the collaboration of Mr. Burnett-Hurst and a number of other workers, Dr. Bowley conducted an inquiry into the conditions of working-class households in four representative towns—Northampton, Warrington, Stanley, and Reading—the results of which are published by Messrs. Bell and Sons under the title of *Livelihood and Poverty*. They are similar in character to those obtained by Rowntree in his study of conditions in York, but what is peculiar to Bowley's inquiry is that only a sample, about 1 in 20, of the working-class houses in each town was examined, and the conditions in the towns as a whole were deduced from these samples.

We are not concerned here with the actual facts disclosed by the investigation, striking as they are, but with the explanation of the sampling method adopted, and as to that it may be remarked that the foundation on which it rests is precisely the same as that which underlay the example of the 999 black and white sheep. The main point to notice here again is that Bowley was careful to select his samples in unbiassed fashion as follows: 'For each town a list of all houses . . . was obtained, and without reference to anything

except the accidental order (alphabetical by streets or otherwise) in the list, one entry in twenty was ticked. The buildings so marked, other than shops, institutions, factories, etc., formed the sample.' It will be evident that this method of choice is not quite on the same level of randomness as that followed, for example, in drawing cards from a well-shuffled pack, each card to be replaced and the pack reshuffled before the next is drawn; but, for that very reason, the results of the experiment are all the more likely to be well within the limits of error provided by the formulæ of the ideal case. The deliberate selection of every twentieth house in each street is likely, that is to say, to give a more representative picture of the town as a whole than would be obtained by selecting the same number of houses in a purely random fashion which might by chance give too much emphasis to some street or district.

A practical test of the goodness of the sample was possible by comparing the results in a few instances with information available from other sources. In order to make the method of working quite clear, let the guiding principle first be recalled :—

'If, in a random sample of n items, the proportion of successes is p, then the *proportion* of successes in the universe from which the sample is selected will not be likely to fall outside the limits

$$p \pm 3(0{\cdot}6745)\sqrt{(pq/n)},$$

and, if that universe contains altogether N items, the *number* of successes will not be likely to fall outside the limits

$$\mathrm{N}p \pm 3(0{\cdot}6745)\mathrm{N}\sqrt{(pq/n)}.\text{'}$$

In Reading the total number of all inhabited houses in the borough was 18,000 at the time of the inquiry, *i.e.* $\mathrm{N}=18{,}000$. The total number of houses visited was 840, *i.e.* $n=840$. If we call a house assessed at £8 or less a 'success,' the number of such houses found in the sample was 206.

Thus $\qquad p=206/840, \qquad q=634/840,$

and the number of houses rented at £8 or less in the whole borough should be

$\mathrm{N}p$ with a p.e.$=0{\cdot}6745\mathrm{N}\sqrt{(pq/n)}$

i.e. $\qquad 4414 \pm 180.$

The actual number of houses so rented was known from other sources to be 4380, well within the limits forecasted.

The value used for p in the above is that given by the sample, but when we know the actual number of successes in the universe

as a whole, as in this case we do, we might use the true value of p, *i.e.* the value for the universe in place of that for the sample. The argument might also be put in another way without affecting the principle employed, thus :—

The number of houses rented at £8 or less in the whole borough was 4380.

But the proportion of houses sampled in the whole borough was 840/18000, *i.e.* 1/21·43.

Hence the number of houses at the above rental to be expected in the sample=4380/21·43=204.

The number actually found in the sample was 206, with a probable error

$$=0{\cdot}6745\sqrt{(npq)}$$
$$=0{\cdot}6745\sqrt{(840\times\tfrac{4380}{18000}\times\tfrac{13620}{18000})}$$
$$=8, \text{ approximately.}$$

Again, the number of persons engaged in a certain occupation at Reading was known to be 761 in the borough as a whole. Hence the number of persons so engaged to be expected in the sample was 761/21·43, *i.e.* 35.

The number actually found in the sample was 29 with a probable error

$$=0{\cdot}6745\sqrt{(npq)}$$
$$=0{\cdot}6745\sqrt{(840\times\tfrac{761}{18000}\times\tfrac{17239}{18000})}$$
$$=4, \text{ approximately.}$$

Further examples of the method are here given, in each of which the total number of events is small so that the number in each sample is also small, and since, as we have seen, the accuracy or precision of the proportion of successes discovered in any sample varies directly as the square root of the number of events the sample contains, the results cannot be expected to be so good when this number is small.

Example (8).—514 candidates sat a certain examination paper; their marks ranged from 3 to 64. The candidates were numbered consecutively from 1 to 514, and a random sample of 90 (17½ per cent.) was selected from among them by writing down the 90 numbers formed by the digits in the seventh decimal place, taken in groups of three, in the logs of the numbers 10104, 10204, 10304, . . . , as given in Chambers's Tables, neglecting all numbers greater than 514 and calling such numbers as 005, 037, etc.—5, 37, etc. In this way each of the numbers between 1 and 514 stood an equal chance of inclusion.

The distribution of candidates in the sample is compared with that for all 514 together in the following table :—

No. of Marks Obtained.	Percentage of All Candidates who obtained these Marks.	Percentage of Candidates in Sample who obtained these Marks.
		p.e.
Less than 15	8	$8\pm1\cdot9$
15 but less than 25	19	$17\pm2\cdot6$
25 „ „ 30	16	$18\pm2\cdot7$
30 „ „ 35	18	$13\pm2\cdot4$
35 „ „ 40	15	$17\pm2\cdot6$
40 „ „ 50	19	$18\pm2\cdot7$
50 and over.	7	$10\pm2\cdot1$

The reader might verify the p.e.'s given in the last column : *e.g.* proportion in the sample obtaining less than 15 marks$=7/90$; therefore $p=7/90, \quad q=83/90.$

Hence the S.D. for this group

$$=\sqrt{[7(1-\tfrac{7}{90})]}$$
$$=2\cdot54,$$

and the S.D. for the percentage

$$=\tfrac{100}{90}\times2\cdot54=2\cdot8.$$

Thus the p.e. for the percentage

$$=\tfrac{2}{3}\sigma=1\cdot9, \text{ approximately.}$$

Example (9) deals in a similar way with the data concerning infectious diseases in 241 towns in England and Wales previously recorded on p. 62.

A sample of 60 towns, *i.e.* about 25 per cent., was chosen in a random fashion as in the last example, and the sample distribution is compared below with that of the 241 towns as a whole.

The verification of the probable errors in this and the next case is left to the reader.

Case Rate per 1000 of the Population.	Actual No. of Towns so rated.	No. as suggested by the Sample.
		p.e.
1 and under 5	85	92 ± 10
5 „ 9	86	96 ± 10
9 „ 13	42	$28\pm\ 7$
13 and over.	28	$24\pm\ 6$

Example (10) is concerned with the annual output per head in 142 different types of employment as given in 1907 by the *Census of Production* [data from *Sixteenth Abstract of Labour Statistics of the United Kingdom,* Cd. 7131]. The distribution suggested by a random sample of 50 different occupations is compared with that of the complete list of 142 occupations.

Output per head.	No. of Occupations in Sample with this Output.	No. in Complete List as deduced from Sample.	Actual No. found in Complete List.
		p.e.	
Under £60 . .	4	11±3·6	12
£60 and under £80	16	45±6·2	42
£80 „ £100	6	17±4·3	25
£100 „ £120	10	28±5·3	20
£120 „ £190	8	23±4·9	27
£190 and over .	6	17±4·3	16

The S.D. in each of the last three examples has been calculated by using the value for p given by the sample, which is the value one must fall back upon in practice when the true p for the whole distribution is unknown. In any case where we are able to test our sample by comparison with the whole distribution, however, it is possible to use the true value of p, *e.g.* in Example (10) output £100-120, $p=20/142$ as opposed to 10/50.

CHAPTER XV

CURVE FITTING—PEARSON'S GENERALIZED PROBABILITY CURVE

It may be recalled that in the introductory chapter an outline was given of the manner in which the theory of Statistics might be conceived to develop. It was shown how the desire for simplification and the need for compression leads to the division of a large mass of figures dealing with any given matter into groups; indeed, it may well be that the statistics have been so arranged at the source in the act of collecting: *e.g.* we may have to deal with so many males of height 54 in. and less than 55 in., so many of height 55 in. and less than 56 in., so many of height 56 in. and less than 57 in., and so on. Here corresponding to each given height, which we may label x, or each range of height, such as x_1 to x_2, we have a certain frequency of males of that height or range, which frequency we may label y, and hence a frequency table can be formed showing the variation of y with x. Further we have seen how such pairs of corresponding values of x and y can be plotted so as to picture the complete *observed* frequency distribution to the eye.

Now the representation thus made, though helpful up to a point, is not entirely satisfactory. Whether we simply join up successive points (x, y), or set up rectangles of varying height y on bases spanning the successive ranges of x, or erect ordinates (y's) at the mid-points of these bases, joining the summits in the manner previously described, the connection so established between each observation and the next is too superficial, depending merely on the fact of casual neighbourship, and may sometimes give a false impression of frequency and changes in frequency in the population of which the observations are but a sample. And this is necessarily so if we confine ourselves strictly to the data observed.

One difficulty which has to be faced is that only within certain broad limits can we trust our observations to give us information which is truly representative of the population in which we are

interested. We seldom if ever deal with the whole population: in fact it may be so large that it is impracticable even to reckon it; instead we make a random or unbiassed selection of a smaller but adequate number of individuals belonging to the population, and classify them according to the size or nature of the character which concerns us. But, granted that our sample is adequate in size and unbiassed, the numbers obtained in the different groups of the frequency distribution will still be subject to the errors of random sampling, and it is only after these errors have been calculated that we can lay down the probable limits within which our sample may be regarded as really representative of the population as a whole.

Another difficulty arises owing to the fact that our observations in general do not cover the whole field of values of the variables x and y; we may quite likely want to know the percentage frequency, y, of individuals with a character (height or whatever it may be) x which does not chance to be any one of the x's observed, if the observations are only recorded according to discrete (separately distinct, like 5 ft., 6 ft., 7 ft.) values of x; on the other hand, if the observations have been classed in groups, the frequency in which we are interested may refer to an x which does not coincide with the centre of any group or which is even outside the range altogether. We have therefore further to inquire whether such information can be deduced in any way from the statistics collected.

Now it so happens that both these difficulties disappear if we can only attain the ideal already outlined in discussing graphs, and find a suitable curve to 'fit' the statistics observed. Such a curve would not necessarily pass through all or any of the points (x, y) representing the observations, for these, as we have remarked, are subject to errors of random sampling and the observed frequency y of any x may be greater or less than the corresponding y in the population at large to which the curve is presumed to approximate. The curve in short must remove the roughnesses which are inseparable from ordinary observation. Moreover, given any x, not merely one of the x's observed, it must be possible to read off from it the corresponding y, the frequency appropriate to that x.

It is not always accurate enough for our purpose to draw a curve by eye, passing as evenly as possible through the middle of the points observed in the manner conceived in an earlier chapter. It is necessary in some way to find an algebraical formula, possibly even a trigonometrical, exponential, or more complex expression, which will give the y corresponding to any x desired. This formula or equation must depend upon the statistics collected: *i.e.* the

constants involved in it must be directly and fairly easily computed from the y's observed, and the results of *all* the observations should enter into the equations which determine the constants in order to make use of the full information at our disposal. In addition, the method of determining the equation and its constants should be as general as possible, so relieving us of the trouble of discovering a new method owing to the failure of the original one at nearly every trial. Finally, the equation should not be so intricate as to make the labour of calculating y for any given x too heavy to be attempted with the ordinary equipment at the statistician's disposal. Once such an equation is found it is a fairly straightforward proceeding to trace the curve for which it stands, and it will remain afterwards to test the goodness of fit in some more refined way than by seeing how closely it passes through the observed points by eye.

When we come to review the shapes of the frequency polygons or histograms most commonly met, we find that the majority of them start from low frequency, rise to a maximum as x, the character observed, increases, then fall again towards zero very likely at a different rate. In fact the statistics suggest a shape something like that shown in fig. (27) for the corresponding frequency curve, though we cannot be sure that it would coincide with the axis at either extremity. [Cases do occur where the curve has two or even more humps (maxima), but we purposely restrict ourselves to the simpler and more frequent type described.]

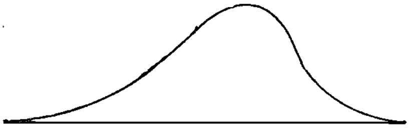


Fig. (27).

Now the simplest shape to deal with from the algebraical point of view would certainly be symmetrical in character, corresponding to statistics which rise and fall at the same rate, though this would not necessarily be the most common shape among the records of actual life. In order to simplify our problem, therefore, we might start by making up for ourselves an ideally simple set of statistics which are perfectly symmetrical, and see whether we can discover a process for fitting a curve in a case of that kind. If this prove successful it might be possible afterwards to adapt the same process to an unsymmetrical or 'skew' set of statistics made up in a similar way. Then finally we should inquire whether actual observations conform to any of the types of curve discovered, and, if so, how they can be fitted together.

Now in manufacturing our statistics we must keep before us the

object at which we are aiming. Given the statistics, what we want is a formula, algebraical or of some other kind, to fit them. This raises the possibility of choosing the statistics themselves in some algebraical form, and such a form is at hand in the binomial expansion, which is, in fact, one of the first examples of a general symmetrical expression one meets. Thus

$$(a+b)^1=a+b$$
$$(a+b)^2=a^2+2ab+b^2$$
$$(a+b)^3=a^3+3a^2b+3ab^2+b^3$$
$$(a+b)^4=a^4+4a^3b+6a^2b^2+4ab^3+b^4$$
$$(a+b)^5=a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$(a+b)^n=a^n+na^{n-1}b+\frac{n(n-1)}{1\cdot 2}a^{n-2}b^2+\ \ldots$$
$$+\frac{n(n-1)}{1\cdot 2}a^2b^{n-2}+nab^{n-1}+b^n.$$

Clearly all these expressions become perfectly symmetrical if we put $a=b$, for they read the same whether we run from left to right or from right to left.

We have already seen what an important part the binomial expansion plays in the early stages of the theory of probability: *e.g.* $(\frac{1}{2}+\frac{1}{2})^{10}$, when expanded, tells us at once the proportion of times on the average we may expect 10 heads, 9 heads and 1 tail, 8 heads and 2 tails, and so on, when we toss an evenly-balanced coin ten times in succession; or again, if p is the probability that a certain event will happen, and q the probability that it will fail to happen at one trial, then the probabilities that it will happen p times, $(p-1)$ times, $(p-2)$ times, . . . in n trials are given by the successive terms in the expansion of $(p+q)^n$. However, we make no assumption for the moment as to the values of a and b, except that in the symmetrical case with which we begin they are equal, and we have as the successive terms of $(a+a)^n$:—

$$a^n,\ na^n,\ \frac{n(n-1)}{1\cdot 2}a^n,\ \ldots\ ,\ \frac{n(n-1)}{1\cdot 2}a^n,\ na^n,\ a^n.$$

Let us suppose that our observed statistics take the above form so that these terms may be plotted, as a succession of ordinates, $y_1, y_2, y_3, \ldots, y_{n+1}$, associated with abscissæ, $x_1, x_2, x_3, \ldots, x_{n+1}$, at equal distances apart measured, say, by c; for convenience we may place the origin as in fig. (28), so that

$$x_1=c,\ x_2=2c,\ x_3=3c,\ \ldots\ ,\ x_{n+1}=(n+1)c,$$

and we can then form a frequency polygon, where

$$x_r = rc,\ y_r = \frac{n(n-1)(n-2)\ .\ .\ .\ (n-r+2)}{1\cdot 2\cdot 3\ .\ .\ .\ (r-1)} a^n,$$

are typical values of a pair of the variables x and y, each such pair defining a vertex of the polygon.

Now in this case, since the statistics have been artificially built up by ourselves and are not in reality a random selection, they are

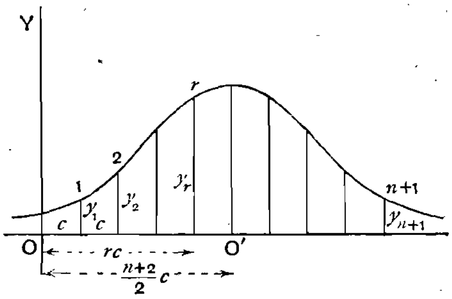
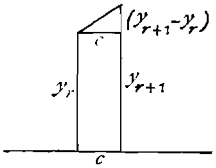


Fig. (28).

not subject to errors of sampling and the fitting curve should, therefore, pass through the summits of all the y's, or, perhaps better, touch each of the lines joining adjacent summits. The curve only differs from the neighbouring outline of the polygon in that the latter is discontinuous, it alters its direction relative to the axis of x by jerks at equal intervals c measured along OX, whereas the former must rise gradually and continuously and then fall in the same way. This is one sense in which we mean that the fitting curve removes the roughness of the observation statistics—it gets rid of jerks besides filling gaps in the observations.

It will be clear that as n increases and c diminishes (and this is what we aim at in collecting statistics, though it has not been assumed in what immediately follows) the discontinuity in the polygon becomes less and less pronounced and the outline of the figure approximates more and more closely to the curve. Moreover this approximation gains in intensity if we make the slope of the curve at each appropriate point the same as the slope obtained by joining up the summits of adjacent ordinates of the polygon.



Now the expression

$$(y_{r+1} - y_r)/c$$

is the measure of the gradient from the rth ordinate to the $(r+1)$th, and

$$\frac{y_{r+1}-y_r}{c}=\frac{a^n}{c}\left[\frac{n(n-1)\ldots(n-r+1)}{1\cdot2\ldots r}-\frac{n(n-1)\ldots(n-r+2)}{1\cdot2\ldots(r-1)}\right]$$

$$=\frac{a^n}{c}\frac{n(n-1)\ldots(n-r+2)}{1\cdot2\ldots(r-1)}\left[\frac{n-r+1}{r}-1\right]$$

$$=y_r\frac{n-2r+1}{rc}.$$

If this be also taken as the gradient of the tangent to the curve at the point midway between (x_r, y_r) and (x_{r+1}, y_{r+1}), calling this point (x, y) we have, since, in the notation of the differential calculus. $\frac{dy}{dx}$ is the measure of the gradient of the curve at this point,

$$\frac{dy}{dx}=\frac{y_{r+1}-y_r}{c}$$

$$=y_r\frac{n-2r+1}{rc}.$$

And

$$x=\tfrac{1}{2}(x_r+x_{r+1})=\tfrac{1}{2}[rc+(r+1)c]=\frac{c}{2}(2r+1)$$

$$y=\tfrac{1}{2}(y_r+y_{r+1})=\frac{a^n}{2}\frac{n(n-1)\ldots(n-r+2)}{1\cdot2\ldots(r-1)}\left[\frac{n-r+1}{r}+1\right]=\frac{y_r}{2r}(n+1).$$

Hence

$$y_r\frac{n-2r+1}{rc}=\frac{2ry}{n+1}\cdot\frac{(n+2)-(2r+1)}{rc}=\frac{2y}{(n+1)c}\left(n+2-\frac{2x}{c}\right).$$

Thus

$$\frac{dy}{dx}=\frac{2y}{(n+1)c}\left(n+2-\frac{2x}{c}\right).$$

But if we had started with any other two adjacent ordinates instead of y_r and y_{r+1} we should have been led to exactly the same relation connecting the corresponding x and y of the required curve, for r, which serves to particularize the ordinates, does not appear in the relation at all—their individuality has been eliminated. The above equation may thus, if we please, be taken as holding good for, and therefore defining, *all* points (x, y) of the fitting curve: it is, in short, the differential equation of that curve.

The equation may be slightly simplified by transferring the origin to the point $\left[(n+2)\frac{c}{2}, 0\right]$, evidently the point O′ in fig. (28)

corresponding to the maximum ordinate of the polygon or curve. Algebraically, this merely means that for x we must write $\left[x+\frac{(n+2)c}{2}\right]$ in the equation, which then becomes

$$\frac{dy}{dx}=\frac{2y}{(n+1)c}\left(-\frac{2x}{c}\right)=-\frac{4xy}{(n+1)c^2}.$$

We may pass to the equation proper of the curve by integration. Thus, separating the variables,

$$\int\frac{dy}{y}+\frac{4}{(n+1)c^2}\int xdx=0.$$

Therefore, $$\log y+\frac{2x^2}{(n+1)c^2}+\mathrm{A}=0,$$

where A is a constant.

Hence $$y=y_0e^{-2x^2/c^2(n+1)},$$

where y_0 is a new constant.

This may be written

$$y=y_0e^{-x^2/2\sigma^2}, \quad . \quad . \quad . \quad (1)$$

where $\sigma^2=(n+1)c^2/4$, and it is called *the probability curve* or *normal curve of error.**

Let us now see whether the procedure so far followed is applicable in the case of an unsymmetrical or skew distribution of statistics. With this object we will suppose the frequencies of observations in successive groups to be represented by the corresponding terms in the expansion

$$(p+q)^n=p^n+np^{n-1}q+\frac{n(n-1)}{1\cdot2}p^{n-2}q^2+\ .\ .\ .\ ,$$

and as before we can form a frequency polygon by joining the summits of the ordinates

$$y_1=p^n,\ y_2=np^{n-1}q,\ y_3=\frac{n(n-1)}{1\cdot2}p^{n-2}q^2,\ .\ .\ .\ ,\ y_{n+1}=q^n,$$

[* Karl Pearson's method of getting the normal curve equation has been adopted as the basis of the above discussion, in preference to that usually followed, which develops the curve also from the binomial expression but somewhat on the lines of Laplace and Poisson. They showed that the sum of all the terms lying within a range t on either side of the maximum term in the expansion of $(p+q)^n$ is approximately

$$=\frac{1}{\sqrt{2\pi}\cdot\sigma}\left[\int_{-t}^{+t}e^{-x^2/2\sigma^2}dx+e^{-t^2/2\sigma^2}\right],$$

where $\sigma=\sqrt{(npq)}$, whence the equation of the curve is derived. (See Historical Note at the end of Chapter XVIII.)]

erected on the axis of x at distances from the origin given by

$$x_1=c,\ x_2=2c,\ x_3=3c,\ \ldots,\ x_{n+1}=(n+1)c,$$

the figure being very similar to that in the symmetrical case.

The gradient of the fitting curve where it touches the join of (x_r, y_r) to (x_{r+1}, y_{r+1}) is given by

$$\frac{dy}{dx}=\frac{y_{r+1}-y_r}{c},$$

and we must try and express the right-hand side as before in terms of (x, y), the co-ordinates of the mid-point of the line joining (x_r, y_r) to (x_{r+1}, y_{r+1}).

We have

$$\frac{dy}{dx}=\frac{1}{c}\left[\frac{n(n-1)\ldots(n-r+1)}{1\cdot2\ldots r}p^{n-r}q^r-\frac{n(n-1)\ldots(n-r+2)}{1\cdot2\ldots(r-1)}p^{n-r+1}q^{r-1}\right]$$

$$=\frac{p^{n-r}q^{r-1}}{c}\cdot\frac{n(n-1)\ldots(n-r+2)}{1\cdot2\ldots(r-1)}\left[\frac{n-r+1}{r}q-p\right].$$

Also

$$2x=x_r+x_{r+1}=rc+(r+1)c=(2r+1)c$$

$$2y=y_r+y_{r+1}=\frac{n(n-1)\ldots(n-r+2)}{1\cdot2\ldots(r-1)}\cdot p^{n-r}q^{r-1}\left[\frac{n-r+1}{r}q+p\right].$$

Thus

$$\frac{dy}{dx}=\frac{2y}{c}\left(\frac{n-r+1}{r}q-p\right)\Big/\left(\frac{n-r+1}{r}q+p\right)$$

$$=\frac{2y}{c}[(n+1)q-r(p+q)]/[(n+1)q+r(p-q)]$$

$$=\frac{2y}{c}[2(n+1)qc-(p+q)(2x-c)]/[2(n+1)qc+(p-q)(2x-c)].$$

This, being true for *all* such pairs of values of x and y, is now in a form independent of *any particular point* on the curve we seek; in other words, it may be taken as the differential equation of the curve, and it is evidently of the type

$$\frac{dy}{dx}=y\frac{(\alpha-x)}{(\beta+\gamma x)},$$

where α, β, γ involve only p, q, n, etc., the constants of the distribution we set out to fit.

The equation is simplified if we transfer the origin to the point $(a, 0)$, when it becomes

$$\frac{dy}{dx}+\frac{yx}{\gamma x+\delta}=0,$$

where $\delta=\beta+\gamma a$.

To integrate, separate the variables as before:

$$\int\frac{dy}{y}+\int\frac{x}{\gamma x+\delta}dx=0.$$

Therefore, $$\log y+\frac{1}{\gamma}\int\frac{(\gamma x+\delta)-\delta}{\gamma x+\delta}dx=0$$

$$\log y+\frac{x}{\gamma}-\frac{\delta}{\gamma^2}\log(\gamma x+\delta)+\mathrm{A}=0,$$

where A is a constant,

or $$y=\mathrm{B}e^{-x/\gamma}(\gamma x+\delta)^{\delta/\gamma^2},$$

where B is a constant.

It may be written

$$y=y_0e^{-kx}\left(1+\frac{x}{a}\right)^{ka} \quad . \quad . \quad . \quad (2)$$

where $k=1/\gamma$, $a=\delta/\gamma$, and y_0 is a new constant.

This, then, may prove a suitable type of curve to fit a set of statistics forming a skew frequency distribution, but the question now arises whether equations (1) and (2) are the most general types possible. Clearly (1) is only a particular case of (2) obtained by making $p=q$, and, this being so, (2) may itself be a particular case of some still more general type.

Light may be thrown on this if we consider the geometrical bearing of the differential equation obtained in the last case:

$$\frac{dy}{dx}=\frac{y(a-x)}{\beta+\gamma x} \quad . \quad . \quad . \quad . \quad (3)$$

The presence of y and $(a-x)$ in the numerator of the right-hand side of (3) shows that $\frac{dy}{dx}$ vanishes when $y=0$ and when $x=a$, *i.e.* the curve touches the axis of x where the two meet and there is a maximum point on the curve at $x=a$. (Since a is the particular value of the organ or character x for which the frequency is a maximum, a is of course the mode.) Now these two characteristics are the very ones to which we wished to give symbolical expression since they serve to describe in broad outline what was agreed to

be the trend of the majority of frequency distributions—the rise from zero to a maximum, at first gradually, then faster, and, after passing through the maximum, the fall to zero again, generally at a different rate.

As to the denominator of equation (3), the corresponding equation for type (1), before the origin was changed, was similar to equation (3), except that it contained no x term in the denominator, and that is readily understood when we note that γ is a multiple of $(p-q)$ and thus vanishes when $p=q$. Now, if from (3) we get a less general type of curve by dropping the x term in the denominator, we may perhaps get a more general type by adding an x^2 term, and even an x^3 term, an x^4 term, and so on. In fact there seems no reason why the denominator should not be any function of x, say $f(x)$, which, however, we shall suppose for simplicity capable of expansion in a Maclaurin's series of ascending powers of x which converges quickly.

We are led to propose, therefore, as more general than (3), the differential equation

$$\frac{dy}{dx}=\frac{y(x+b)}{px^2+qx+r} \quad . \quad . \quad . \quad (4)$$

We stop at x^2 in the denominator because it has been found, if we may anticipate results to save needless labour, that beyond this point the heaviness of the calculation involved and the decreasing accuracy of the higher moments that have to be introduced outweigh any other advantage gained. The curve or set of curves resulting from the integration of equation (4) is known as Karl Pearson's Generalized Probability Curve, and their author has stated that, while it comprises the two other types as special cases, it practically covers all homogeneous statistics he has had to deal with.

Just as the differential equations in the first two cases considered were related respectively to the symmetrical and the skew binomial expansions, so is equation (4) related to the hypergeometrical expansion

$$[{}^{pn}C_r+{}^{pn}C_{r-1}{}^{qn}C_1+{}^{pn}C_{r-2}{}^{qn}C_2+\ .\ .\ .\ +{}^{qn}C_r]/{}^{n}C_r,$$

the successive terms of which express the probability that r black balls, $(r-1)$ black balls and 1 white ball, $(r-2)$ black balls and 2 white balls, . . ., r white balls, will be drawn from a bag containing pn black balls and qn white ones, where $(p+q)=1$, when r balls are drawn in all, each being replaced before the next is drawn.

If the terms of this expansion are represented by ordinates of which the summits determine a polygon as in the binomial cases, the corresponding expression for the gradient of the curve at any point is given by an equation of type (4). We need not go over the detailed proof of this statement since it follows precisely the same lines as in the previous cases.

The method of integration of the equation

$$\frac{dy}{dx}=\frac{y(x+b)}{px^2+qx+r}$$

depends upon the nature of the roots of the quadratic in the denominator which may be written

$$px^2+qx+r=p\left[\left(x+\frac{q}{2p}\right)^2-\left(\frac{q^2}{4p^2}-\frac{r}{p}\right)\right]$$

$$=p\left[\left(x+\frac{q}{2p}\right)^2-\frac{4r^2}{q^2}\cdot\frac{q^2}{4pr}\cdot\left(\frac{q^2}{4pr}-1\right)\right]$$

$$=p\left[\left(x+\frac{q}{2p}\right)^2-\frac{4r^2}{q^2}\kappa(\kappa-1)\right],$$

where $\kappa=q^2/4pr$, and it is evident that the quadratic splits up into real factors if $\kappa(\kappa-1)$ is positive. This is the case when κ has any negative value, or when it is positive and greater than 1, the truth of which may be seen more effectively if the curve

$$y=\kappa(\kappa-1),$$

a parabola symmetrical about the line $\kappa=\frac{1}{2}$, be drawn, fig (29), by plotting y against κ.

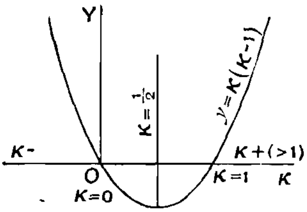


Fig. (29).

Further, the product of the roots of the quadratic

$$px^2+qx+r=0$$

is

$$\frac{r}{p}=\frac{4r^2}{q^2}\cdot\frac{q^2}{4pr}=\frac{4r^2}{q^2}\cdot\kappa,$$

so that the roots when real will be of the same sign if κ is positive and of opposite signs if κ is negative. The boundary lines

$$\kappa=0 \text{ and } \kappa=1$$

thus divide the whole field into three parts, as shown in fig. (30), in one of which the roots are real and of opposite sign, in the next

the roots are imaginary, and in the third the roots are real and of the same sign. At the boundaries we get particular cases as follows :—

$\kappa=0$: this requires $q=0$, since $\kappa=q^2/4pr$, which makes the roots of the quadratic equal but of opposite sign, unless $p=0$ also, and in that case both roots are infinite ;

$\kappa=1$: the roots are real and equal and of the same sign ;

$\kappa=\infty$: this requires $p=0$ or $r=0$; in the former case one root of the quadratic is infinite, and in the latter one root is zero.

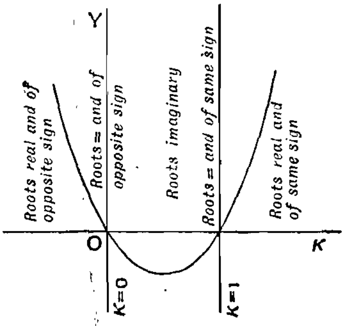


FIG. (30).

Thus, returning to the differential equation, the curves which result from the integration

$$\int\frac{dy}{y}=\int\frac{(x+b)dx}{px^2+qx+r}$$

are of different types according to the value of κ, which is therefore called the *criterion.*

Type I.—κ—ve. *Roots of* $px^2+qx+r=0$ *real and of opposite sign.* In this case we may write

$$px^2+qx+r=p(x+a')(x-\beta')$$

and so get

$$\int\frac{dy}{y}+\int\frac{(x+b)dx}{p(a'+x)(\beta'-x)}=0,$$

or, transferring the origin to the point $(-b,\ 0)$, the mode, we have

$$\int\frac{dy}{y}+\int\frac{xdx}{p(a'-b+x)(\beta'+b-x)}=0,$$

or
$$\int\frac{dy}{y}+\int\frac{xdx}{p(a+x)(\beta-x)}=0,$$

where $a=a'-b,\ \beta=\beta'+b.$

Therefore, $$\log y-\frac{1}{p}\int\frac{a}{a+x}\,\frac{dx}{a+\beta}+\frac{1}{p}\int\frac{\beta}{\beta-x}\,\frac{dx}{a+\beta}+\mathrm{A}=0,$$

where A is a constant.

Thus $$\log y = \frac{1}{p(a+\beta)}[a \log (a+x) + \beta \log (\beta - x)] + \log \mathrm{B},$$

where B is a constant,

whence $$y = \mathrm{B}(a+x)^{\frac{a}{p(a+\beta)}}(\beta - x)^{\frac{\beta}{p(a+\beta)}}$$

i.e. $$y = y_0\left(1+\frac{x}{a}\right)^{\nu a}\left(1-\frac{x}{\beta}\right)^{\nu\beta} \quad . \quad . \quad . \quad (5)$$

where $\nu = 1/p(a+\beta)$ and y_0 is a new constant.

This is a skew curve of limited range, bounded by the lines $x=-a$ and $x=+\beta$, with the mode at the origin.

Type II.—$\kappa=0$. *$q=0$, but not $p=0$. Roots of $px^2+qx+r=0$ equal and of opposite sign.*

This curve is just a particular case of type I., which reduces to

$$y = y_0\left(1-\frac{x^2}{a^2}\right)^{\nu a}, \quad . \quad . \quad . \quad (6)$$

symmetrical about the axis of y (because for any value of y there are two values of x, equal and of opposite sign) and of limited range bounded by $x=-a$ and $x=+a$, with the mode at the origin.

Type III.—$\kappa=\infty$.* *$p=0$, but not $r=0$. One root of $px^2+qx+r=0$ infinite.*

This is the skew binomial case over again. It may be also deduced from type I. by making one root, say β', tend to infinity. The curve then takes the form

$$y = y_0\left(1+\frac{x}{a}\right)^{\nu a} \underset{\beta\to\infty}{L}\left(1-\frac{x}{\beta}\right)^{\nu\beta},$$

because $\boldsymbol{\beta=\beta'+b}$, so that β tends to infinity with β'. Hence

$$y = y_0\left(1+\frac{x}{a}\right)^{\nu a} \underset{\lambda\to\infty}{L}\left[\left(1+\frac{1}{\lambda}\right)^{\lambda}\right]^{-\nu x},$$

where $\boldsymbol{\lambda = -\beta/x}$.

Thus $$y = y_0\left(1+\frac{x}{a}\right)^{\nu a} e^{-\nu x}, \quad . \quad . \quad . \quad (7)$$

a skew curve limited in one direction by the line $x=-a$, with the mode at the origin.

[* Although theoretically this type corresponds to an infinite value for κ, in practice it will as a rule give a reasonable fit provided κ is numerically greater than 4. (See W. P. Elderton's *Frequency Curves and Correlation*, p. 50)].

Type IV.—$\kappa +^{ve}$ *and* <1. *Roots of* $px^2+qx+r=0$ *imaginary.*

Put $\kappa(\kappa-1)=-\lambda^2$, and the differential equation then leads to

$$\int\frac{dy}{y}=\int\frac{(x+b)dx}{p\left[\left(x+\frac{q}{2p}\right)^2+\frac{4r^2}{q^2}\lambda^2\right]}.$$

Transfer the origin to the point $\left(-\frac{q}{2p}, 0\right)$

$$\int\frac{dy}{y}=\int\frac{\left(x+b-\frac{q}{2p}\right)dx}{p\left(x^2+4\frac{r^2\lambda^2}{q^2}\right)}$$

$$\log y=A+\frac{1}{2p}\log\left(x^2+4\frac{r^2\lambda^2}{q^2}\right)+\left(\frac{b}{p}-\frac{q}{2p^2}\right)\frac{q}{2r\lambda}\tan^{-1}\frac{xq}{2r\lambda},$$

where A is a constant.

Therefore, $$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}e^{-\nu\tan^{-1}\frac{x}{a}} \quad . \quad . \quad . \quad (8$$

where $$a=\frac{2r\lambda}{q},\ m=-\frac{1}{2p},\ \nu=-\frac{1}{ap}\left(b-\frac{q}{2p}\right),$$

and y_0 is a constant.

This is a skew curve of unlimited range in both directions. The position of the mode is found by putting $\frac{dy}{dx}=0$ in (8) after differentiation, or, what comes to the same thing, is seen by direct reference to the differential equation itself. Thus the distance of the mode from the origin

$$=-\left(b-\frac{q}{2p}\right)=\nu pa$$

$$=-\nu a/2m.$$

Type V.—$\kappa=1$. *Roots of* $px^2+qx+r=0$ ***real and equal.***

The equation to integrate becomes

$$\int\frac{dy}{y}=\int\frac{(x+b)dx}{p\left(x+\frac{q}{2p}\right)^2}.$$

Transfer the origin to the point $\left(-\frac{q}{2p}, 0\right)$, and this becomes

$$\int \frac{dy}{y} = \int \frac{\left(x+b-\frac{q}{2p}\right)}{px^2} dx$$

$$\log y = A + \frac{1}{p} \log x - \frac{1}{p} \cdot \left(b - \frac{q}{2p}\right) \frac{1}{x},$$

where A is a constant.

Therefore,

$$y = y_0 x^{1/p} e^{-\frac{1}{p}\left(b-\frac{q}{2p}\right)\frac{1}{x}}$$

$$\mathbf{y} = \mathbf{y_0 x^{-s} e^{-\gamma/x}}, \quad . \quad . \quad . \quad (9)$$

where $s = -1/p$, $\gamma = \frac{1}{p}\left(b - \frac{q}{2p}\right)$, and y_0 is a constant.

Here x cannot become negative, so that the curve is skew and limited in one direction. The distance of the mode from the origin

$$= -\left(b - \frac{q}{2p}\right) = -p\gamma = \gamma/s.$$

Type VI.—$\kappa + ^{\infty}$ *and* >1. *Roots of* $px^2 + qx + r = 0$ *real and of the same sign.*

Equation becomes

$$\int \frac{dy}{y} = \int \frac{(x+b)dx}{p(x+\alpha)(x+\beta)}$$

$$\log y = \int \left[\frac{b-\alpha}{p(\beta-\alpha)} \cdot \frac{1}{x+\alpha} + \frac{(b-\beta)}{p(\alpha-\beta)} \cdot \frac{1}{x+\beta}\right] dx$$

$$= A + \frac{1}{p(\beta-\alpha)}[(b-\alpha)\log(x+\alpha) - (b-\beta)\log(x+\beta)],$$

where A is a constant;

or, transferring the origin to $(-\beta, 0)$,

$$\log y = A + \frac{1}{p(\beta-\alpha)}\left[\log \{x-(\beta-\alpha)\}^{b-\alpha} - \log x^{b-\beta}\right],$$

$$y = y_0 \{x - (\beta-\alpha)\}^{\frac{b-\alpha}{p(\beta-\alpha)}} x^{-\frac{b-\beta}{p(\beta-\alpha)}}$$

$$\mathbf{y} = \mathbf{y_0 (x - a)^{q_2} x^{-q_1}}, \quad . \quad . \quad . \quad (10)$$

where $a = \beta - \alpha$, $q_2 = (b-\alpha)/p(\beta-\alpha)$, $q_1 = (b-\beta)/p(\beta-\alpha)$, and y_0 is a constant.

This is a skew curve bounded by $x \pm a$ in one direction. The distance of the mode from the origin $= -(b-\beta) = aq_1/(q_1-q_2)$.

Type VII.—$\kappa=0$, $q=0$, $p=0$. *Roots of the quadratic* $px^2+qx+r=0$ *both infinite.*

This is the symmetrical binomial case over again and the integration reduces to

$$\int \frac{dy}{y} = \int \frac{x+b}{r} dx,$$

or, transferring the origin to $(-b, 0)$,

$$\int \frac{dy}{y} = \int \frac{x}{r} dx$$

$$\log y = A + \frac{1}{2r} x^2,$$

where A is a constant.

Therefore $$y = y_0 e^{-x^2/2\sigma^2}, \quad . \quad . \quad . \quad (11)$$

where y_0 is a constant and $\sigma^2 = -r$.

This curve, the normal curve of error, is symmetrical about the axis of y, where mean and mode coincide, and it is of unlimited range on either side of it.

CHAPTER XVI

CURVE FITTING (*continued*)—THE METHOD OF MOMENTS FOR CONNECTING CURVE AND STATISTICS

WE have now completed the first stage of the discussion upon which we embarked : we have found by the application of general principles various types of curve, represented by different equations, which are said to fit more or less satisfactorily a considerable number at all events of frequency distributions composed of homogeneous material.

Our next task is to pass from the general to the particular, to see how to set up a connection between an actually observed frequency distribution and the appropriate theoretical curve. This again seems to break up into two parts—(1) to find a way of deciding which type of curve to adopt in a particular case ; (2) to determine the constants of the curve in terms of the observed statistics ; but since the criterion, κ, which distinguishes one type of curve from another is itself a function of the constants of the curve before integration, it follows that the solution of the first part is incidental to that of the second.

The general method proposed for determination of the constants of the curve in terms of the observed statistics is the now well-known *method of moments* due to Karl Pearson, whereby the area and moments of the fitting curve are equated to the area and moments, calculated from the statistics, of the observation curve.

If a frequency table be drawn up (see Table (40)) showing the number f of observations corresponding to the deviation x of each value, or group mid-value, X of the character observed from some fixed value, the expression

$$x_1f_1+x_2f_2+\ .\ .\ .\ +x_rf_r+\ .\ .\ .$$

is called the first moment of the distribution with reference to the fixed value, which may be termed the origin. Similarly,

$$x^2{}_1f_1+x^2{}_2f_2+\ .\ .\ .\ +x^2{}_rf_r+\ .\ .\ .$$

is called the second moment, Σx^3f, the third moment, Σx^4f, the

fourth moment, and so on. The following notation will be found convenient for working purposes :—

$$N = \Sigma f,\ N'_1 = \Sigma xf,\ N'_2 = \Sigma x^2 f,\ \ldots;$$

$$\nu'_1 = \frac{N'_1}{N} = \frac{\Sigma xf}{\Sigma f},\ \nu'_2 = \frac{N'_2}{N} = \frac{\Sigma x^2 f}{\Sigma f},\ \ldots$$

Undashed letters are reserved for use when the distribution is referred to its mean as origin, in other words when the deviations of the X's are measured from the mean X.

TABLE (40).

Deviation.	Frequency.	First Moment.	Second Moment.	Third Moment.	Fourth Moment.
x_1	f_1	x_1f_1	$x^2{}_1f_1$	$x^3{}_1f_1$	$x^4{}_1f_1$
x_2	f_2	x_2f_2	$x^2{}_2f_2$	$x^3{}_2f_2$	$x^4{}_2f_2$
..	..	..	..	..	..
..	..	..	..	..	..
..	..	..	..	..	..
x_r	f_r	x_rf	$x^2{}_rf_r$	$x^3{}_rf_r$	$x^4{}_rf_r$
..	..	..	..	..	..
..	..	..	..	..	..
Totals .	N	N'_1	N'_2	N'_3	N'_4

Now each N in the frequency table is the sum of a number of discrete quantities which only *tend* to form a continuous series as the class intervals are made very small and the number of observations is made very large. The corresponding frequency polygon or histogram, if we drew it, would at the same time tend to become a continuous curve, the observation curve. If that limiting stage were attainable, if we could actually get an infinitely large sample of observations in which the character observed changed by infinitesimal amounts, we could then replace the isolated f's of observation by the corresponding y''s, the ordinates of this observation curve, and to get the moments we could write instead of the discrete sums

$$\Sigma f,\ \Sigma xf,\ \Sigma x^2 f,\ \ldots,$$

the continuous integral expressions

$$\int y'dx,\ \int xy'dx,\ \int x^2y'dx,\ \ldots,$$

taking in the whole sweep of the curve by integrating throughout

the range of deviation x. We should then have, if areas and moments are equated according to Pearson's method,

$$\int ydx=\int y'dx, \int xydx=\int xy'dx, \int x^2ydx=\int x^2y'dx, \ldots, \int x^nydx=\int x^ny'dx,$$

where y is the ordinate of the fitting curve corresponding to the ordinate y' of the observation curve.

In practice, however, it is impossible to go to this limit: we cannot deal with an infinitely large sample, so we take as large a sample as is convenient, calculate the rough moments, N, N'_1, $N'_2 \ldots$, and find approximately what corrections or adjustments are necessary to obtain the moments of the observation curve, a procedure which is really equivalent to the determination of the area of a curve when only a number of isolated points thereon are known.

For the full analytical justification of the method of moments the reader is referred to Professor Pearson's original paper, *On the Systematic Fitting of Curves to Observations and Measurements* [*Biometrika*, vol. i., pp. 265 *et seq.*; also vol. ii., pp. 1–23], where it is shown that 'with due precautions as to quadrature, it gives, when one can make a comparison, sensibly as good results as the method of least squares.' The latter, which is the traditional way of approaching all such problems, is shown to be impracticable in a large number of cases, either because the resulting equations cannot be solved, or, when they are capable of solution, because the labour involved would be colossal.

Let us consider next how to deduce the area and moments of the observation curve from the statistics, in other words how to get

$$\int y'dx, \int xy'dx, \int x^2y'dx, \ldots,$$

the integrals being taken throughout the range of the curve, when we know the frequencies corresponding to only a certain number of values or elementary ranges of the deviation x.

Now the character observed may be capable of the deviations actually recorded and of no values in between, *e.g.* measuring deviations from 'no rooms' as origin, we might have f_1 one-roomed tenements, f_2 two-roomed tenements, f_3 three-roomed tenements, but there could be no such thing as a two-and-a-half or a three-and-a-quarter-roomed tenement; on the other hand, any recorded deviation, x_r, may be only the mid-value (used as a convenient and concise approximation) of a group of observations including all in the continuous range from $(x_r-\frac{1}{2})$ to $(x_r+\frac{1}{2})$, where unit deviation is the class interval: thus we might have f_1 males deviating by +6 in. from 5 ft. (comprising all the males observed between 5 ft.

$5\frac{1}{2}$ in. and 5 ft. $6\frac{1}{2}$ in.), f_2 males deviating by $+5$ in. from 5 ft. (comprising all males between 5 ft. $4\frac{1}{2}$ in. and 5 ft. $5\frac{1}{2}$ in.), and so on. These two cases must be discussed separately.

(1) *When the observations are centred at definite but isolated values of x.*

The problem is to find

$$\int x^n y' dx$$

(the nth moment) when we have no definite curve given but we know the values of x and y' at a number of isolated points, say

$$(x_0, y'_0), (x_1, y'_1), (x_2, y'_2), \ldots, (x_p, y'_p).$$

This is equivalent to discovering a suitable 'quadrature formula,' *i.e.* a good approximation to

$$\int z dx$$

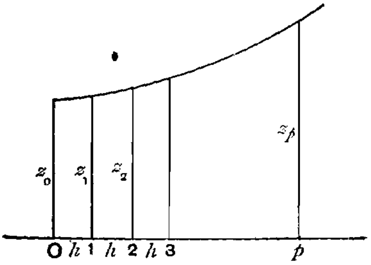


Fig. (31).

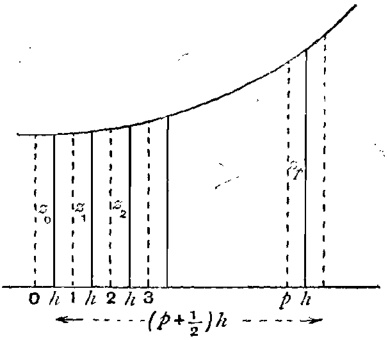


Fig. (32).

in terms of known points

$$(x_0, z_0), (x_1, z_1), (x_2, z_2), \ldots (x_p, z_p),$$

where we have written z in place of $x^n y'$, and we may generally take the ordinates to be at equal distances, h, apart. Several such formulæ have been suggested and they vary according as the z's are situated at the ends (fig. (31)) or at the centres (fig. (32)) of the h intervals. The second type is perhaps the more useful of the two, and we shall work out one formula in illustration of it.

Consider the first five of the given points, namely,

$$(x_0, z_0), (x_1, z_1), \ldots (x_4, z_4).$$

As a simple 'curve of closest contact' let us find the parabola of type

$$z = c_0 + c_1 x/h + c_2 x^2/h^2 + c_3 x^3/h^3 + c_4 x^4/h^4 \quad . \quad . \quad (1)$$

which goes through these five points, where the c's are constants to be determined. We may without loss of generality take the axis

of z to coincide with the middle one of the five ordinates, so that the known points on the curve become

$$(-2h,\ z_0),\ (-h,\ z_1),\ (o,\ z_2),\ (+h,\ z_3),\ (+2h,\ z_4),$$

and on substitution in (1) we get

$$z_0=c_0-2c_1+4c_2-8c_3+16c_4. \qquad z_1=c_0-c_1+c_2-c_3+c_4.$$
$$z_2=c_0. \qquad z_3=c_0+c_1+c_2+c_3+c_4.$$
$$z_4=c_0+2c_1+4c_2+8c_3+16c_4.$$

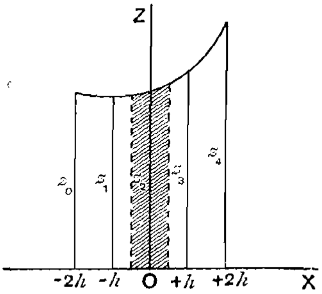


Fig. (33).

These equations are just sufficient uniquely to determine the c's, and hence the parabolic curve of closest contact, in terms of the five given points, but for our purpose it is not necessary to find all the c's. Suppose our object is to find the area of the shaded portion of fig. (33) in terms of the co-ordinates of the five given points. This area

$$=\int_{-h/2}^{+h/2} z\,dx$$
$$=\int_{-h/2}^{+h/2} (c_0+c_1x/h+c_2x^2/h^2+c_3x^3/h^3+c_4x^4/h^4)dx$$
$$=\Big[c_0x+c_1x^2/2h+c_2x^3/3h^2+c_3x^4/4h^3+c_4x^5/5h^4\Big]_{-h/2}^{+h/2}$$
$$=c_0h+c_2h/12+c_4h/80.$$

But the equations between the z's and c's at once give

$$z_2=c_0,\ z_0+z_4=2(c_0+4c_2+16c_4),\ z_1+z_3=2(c_0+c_2+c_4).$$

Thus

$$\left.\begin{aligned}8c_2+32c_4&=(z_0+z_4)-2z_2\\ 2c_2+2c_4&=(z_1+z_3)-2z_2\end{aligned}\right\}.$$

Therefore

$$24c_2=16(z_1+z_3)-(z_0+z_4)-30z_2$$
$$24c_4=(z_0+z_4)-4(z_1+z_3)+6z_2.$$

Hence, by substitution, the shaded area becomes

$$\int_{-h/2}^{+h/2} z\,dx=h\Big[z_2+\tfrac{1}{288}\{16(z_1+z_3)-(z_0+z_4)-30z_2\}$$
$$+\tfrac{1}{1920}\{(z_0+z_4)-4(z_1+z_3)+6z_2\}\Big]$$
$$=\frac{h}{5760}\big[5178z_2-17(z_0+z_4)+308(z_1+z_3)\big], \quad \text{(2)}$$

these particular ordinates being appropriate when the axis of z coincides with the z_2 ordinate.

Similarly, it can be shown that

$$\int_{-h/2}^{+3h/2} z\,dx = \frac{h}{24}[27z_0 + 17z_1 + 5z_2 - z_3], \quad . \quad . \quad . \quad . \quad . \quad (3)$$

by finding the parabolic curve of closest contact through $(0, z_0)$, (h, z_1), $(2h, z_2)$, $(3h, z_3)$, the axis of z coinciding now with z_0.

Now we require $\displaystyle\int_{-h/2}^{+(p+\frac{1}{2})h} z\,dx$

(see fig. (32)), and this may be obtained by splitting up the integral thus

$$\int_{-h/2}^{+3h/2} + \int_{3h/2}^{5h/2} + \int_{5h/2}^{7h/2} + \ . \ . \ . \ + \int_{(p-\frac{5}{2})h}^{(p-\frac{3}{2})h} + \int_{(p-\frac{3}{2})h}^{(p+\frac{1}{2})h},$$

and applying the formulæ (2) and (3) to evaluate these sub-integrals. The first and last come under head (3), while all the rest come under (2). In fact, we fit together portions of curves of parabolic type based on the successive groups of points

(0, 1, 2, 3), (0, 1, 2, 3, 4), (1, 2, 3, 4, 5), (2, 3, 4, 5, 6), . . .
$(p-4, p-3, p-2, p-1, p)$, $(p-3, p-2, p-1, p)$,

and as the points overlap, in the sense that neighbouring groups have points in common, the curves dovetail into one another and so provide a fairly good approximation to what we want in the way of integral expressions giving areas based upon the positions of certain known points.

We have, then :—

$$\int_{-h/2}^{3h/2} z\,dx = \frac{h}{24}[27z_0 + 17z_1 + 5z_2 - z_3]$$

$$\int_{3h/2}^{5h/2} z\,dx = \frac{h}{5760}[5178z_2 - 17(z_0 + z_4) + 308(z_1 + z_3)]$$

$$\int_{5h/2}^{7h/2} z\,dx = \frac{h}{5760}[5178z_3 - 17(z_1 + z_5) + 308(z_2 + z_4)]$$

$$\int_{7h/2}^{9h/2} z\,dx = \frac{h}{5760}[5178z_4 - 17(z_2 + z_6) + 308(z_3 + z_5)]$$

. . . .

$$\int_{(p-\frac{5}{2})h}^{(p-\frac{3}{2})h} z\,dx = \frac{h}{5760}[5178z_{p-2} - 17(z_{p-4} + z_p) + 308(z_{p-3} + z_{p-1})]$$

$$\int_{(p-\frac{3}{2})h}^{(p+\frac{1}{2})h} z\,dx = \frac{h}{24}[27z_p + 17z_{p-1} + 5z_{p-2} - z_{p-3}].$$

Hence, by addition,

$$\int_{-h/2}^{(p+\frac{1}{2})h} zdx = \frac{h}{5760}[6463z_0+4371z_1+6669z_2+5537z_3+6463z_p$$
$$+4371z_{p-1}+6669z_{p-2}+5537z_{p-3}]$$
$$+h[z_4+z_5+z_6+ \ . \ . \ . \ +z_{p-5}+z_{p-4}]$$
$$=h[1{\cdot}1220(z_0+z_p)+0{\cdot}7588(z_1+z_{p-1})+1{\cdot}1578(z_2+z_{p-2})$$
$$+0{\cdot}9613(z_3+z_{p-3})+(z_4+z_5+ \ . \ . \ . \ +z_{p-4})].$$

In effect, since $z=x^n y'$, this means that to calculate the moments from the given statistics we may work simply with the observed ordinates or frequencies, as drawn up in Table (40), so long as we modify the first four and the last four by multiplying them by suitable factors. In particular, when the frequencies at the beginning and end of the distribution are very small, that is to say, when there is high contact at each end of the frequency curve, we may dispense even with the modifying factors also since we may assume that before the first and after the last ordinate observed there are others which are so small as to be negligible.

Thus, given high contact at each extremity of the observation curve, we may write

$$\int_{-h/2}^{(p+\frac{1}{2})h} zdx = h\Sigma z,$$

or, *if we take the class interval as unit in* ***measuring x so that h*=1,** this gives

$$\int yx^n dx = \Sigma fx^n,$$

where the integral may now be taken as referring to the fitted curve, since the moments of the theoretical and of the observational curves are to be equal, and the integration traverses the extent of the curve. When, however, there is not high contact at the extremities the same equation holds good if we multiply the first and last of the observed f's by 1·1220, the second and the last but one by 0·7588, the third and last but two by 1·1578, and the fourth and last but three by 0·9613.

In particular, when $n=0$, integrating throughout the curve,

$$\int ydx = \Sigma f = \text{N}, \quad . \quad . \quad . \quad (4)$$

which, being interpreted, means that the area contained between the fitting curve and the axis of x measures the total frequency of observations, modified if necessary.

Also, when the observation moments have been adjusted, if we

write μ and μ' in place of ν and ν' in the notation previously proposed (see Table (40)), integrating again throughout the curve,

$$\int xydx/\int ydx = \Sigma xf/\mathrm{N} = \mu'_1, \quad . \quad . \quad . \quad (5)$$

and the geometrical interpretation of this is that the foot of the ordinate passing through the centre of gravity of the area between the fitting curve and the axis registers the deviation of the mean X from the fixed origin.

If deviations are measured from the mean of the distribution as origin $\Sigma(xf)$ vanishes (see also Appendix, Note (5)) so that $\mu_1=0$.

Generally, we have, with the same limits of integration,

$$\int x^n ydx/\int ydx = \Sigma x^n f/\mathrm{N} = \mu'_n,$$

and when the distribution is referred to its mean as origin the right-hand side is written μ_n.

We now pass to the second case.

(2) *When the observations appear in groups ranging between definite values of x, the range of each group as a rule being the same in extent.*

Since the usual procedure here is to treat each member of a group as though it were centred at the x at the middle of that group—*e.g.* a group of school girls each of some weight between 7 stone and 7 stone 5 lbs. would be treated as if all its members were of weight 7 stone 2·5 lbs.—this case evidently reduces to that already considered. It is necessary, however, to examine what correction must be made for *assuming* that all the members of the same group have the same x.

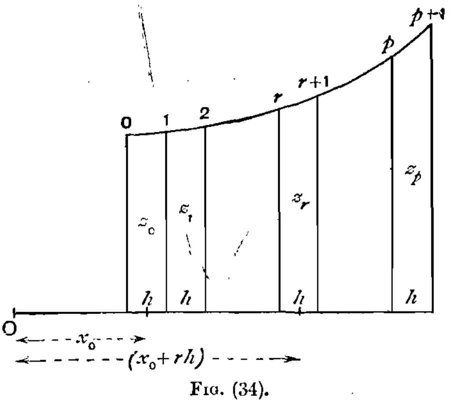


FIG. (34).

Consider again the expression

$$\int x^n y' dx.$$

The contribution to the nth moment coming from the z_r group of observations (see fig. (34)) may be taken as the portion of the above integral between limits $\left(x_0+rh-\frac{h}{2}\right)$ and $\left(x_0+rh+\frac{h}{2}\right)$ where

x_0 is the distance of the centre of the first group from the origin O.

But, since all the observations in the same group are treated as if they had the same x, by (2) this integral may be written

$$\frac{hf_r}{5760}[5178(x_0+rh)^n-17\{(x_0+\overline{r-2}h)^n+(x_0+\overline{r+2}h)^n\}$$
$$+308\{(x_0+\overline{r-1}h)^n+(x_0+\overline{r+1}h)^n\}],$$

where f_r is the frequency of observations in the group, and this, on expansion in powers of (x_0+rh) and h,

$$=hf_r(x_0+rh)^n+\frac{hf_r}{5760}[240n(n-1)h^2(x_0+rh)^{n-2}$$
$$+3n(n-1)(n-2)(n-3)h^4(x_0+rh)^{n-4}+\ . \ . \ .].$$

When we sum *for all groups*, the expression

$$\sum_{r=0}^{r=p} hf_r(x_0+rh)^n$$

gives evidently the nth moment of a set of isolated variables, $f_0, f_1, f_2, \ . \ . \ . \ f_p$, and by Case (1) it may therefore be taken as being practically equivalent to the required nth *moment* of the observation curve, *assuming that there is high contact at each end of the curve.*

The remaining terms,

$$\sum_{r=0}^{r=p}\frac{hf_r}{5760}[240n(n-1)h^2(x_0+rh)^{n-2}$$
$$+3n(n-1)(n-2)(n-3)h^4(x_0+rh)^{n-4}+\ . \ . \ .],$$

may accordingly be taken as the correction required.

When $n=0$, these terms vanish, so we infer, just as in Case (1), that, when the integration is taken throughout the curve,

$$\int ydx=\Sigma f=\mathrm{N}, \quad . \quad . \quad . \quad (4) \text{ bis},$$

or, the area between the fitting curve and the axis of x measures the total frequency of observations when the class interval h is treated as the unit in measuring x.

Again, when $n=1$, the corrective terms vanish, so we likewise infer, as in Case (1), that, with the same limits of integration,

$$\int xydx/\int ydx=\Sigma xf/N=\mu'_1, \quad . \quad . \quad . \quad (5) \text{ bis},$$

and that $\mu_1=0$.

When $n=2$, the reduction of the corrective terms gives

$$\text{second unadjusted moment}=\text{second adjusted moment}+\frac{h^2}{12}\Sigma hf,$$

or, dividing throughout by Σhf and bearing in mind the notation adopted with the mean as origin,

$$\mu_2 = \nu_2 - \tfrac{1}{12}, \quad . \quad . \quad . \quad (6)$$

when $h=1$ as before.

When $n=3$,

third unadjusted moment $=$ third adjusted moment $+\frac{h^3}{4}\Sigma f_r(x_0+rh)$;

but, if we refer the deviations to the mean of the distribution as origin, $\Sigma f_r(x_0+hr)$ vanishes.

Therefore, $$\mu_3 = \nu_3 \quad . \quad . \quad . \quad (7)$$

When $n=4$,

fourth unadjusted moment

$$= \text{fourth adjusted moment} + \frac{h^3}{2}\Sigma f_r(x_0+rh)^2 + \frac{h^4}{80}\Sigma hf.$$

Hence, dividing through as before by Σhf and taking h as 1,

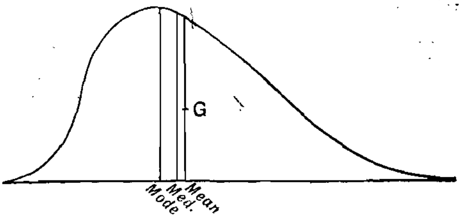
$$\nu_4 = \mu_4 + \tfrac{1}{2}\mu_2 + \tfrac{1}{80}.$$

Therefore, $$\mu_4 = \nu_4 - \tfrac{1}{2}\nu_2 + \tfrac{7}{240} \quad . \quad . \quad . \quad (8)$$

To sum up, the general procedure in Case (2) is to calculate N, N'_1, N'_2, N'_3, N'_4 directly from the statistics and so deduce ν'_1, ν'_2, ν'_3, ν'_4. Then, transferring the origin to the mean, the ν''s become ν_1, ν_2, ν_3, ν_4 (see Appendix, Note 5), and finally the corrected μ's are given by

$$\mu_1 = 0, \qquad \mu_2 = \nu_2 - \tfrac{1}{12},$$
$$\mu_3 = \nu_3, \qquad \mu_4 = \nu_4 - \tfrac{1}{2}\nu_2 + \tfrac{7}{240}.$$

These adjustments, originally due to Dr. W. F. Sheppard * [*Proceedings of the Lond. Mathl. Socy.*, vol. xxix., pp. 353 *et seq.*], are applicable only when the curve of distribution has high contact at each extremity as very frequently happens. To this case we shall confine ourselves, and when it does not hold the unadjusted moments



may be used as a rough approximation failing a more refined but also a more intricate adjustment.

The way in which the three chief kinds of average are related to

[* To obtain Sheppard's adjustments we have followed the method indicated in Elderton's *Frequency Curves and Correlation*, pp. 28, 29.]

the fitting curve is of interest and deserves recapitulation. Whether the observations are classed as in Case (1) or as in Case (2) :—

(1) the ordinate drawn through the highest point of the curve, since the frequency there is a maximum, fixes the *modal value* of X ;
(2) the *median* X is determined by the ordinate bisecting the area between the curve and axis, since there are an equal number of observations on either side of it ; and
(3) the *mean* is determined by the ordinate through the centre of gravity of the area between the curve and axis.

We have still to show how to express the constants of the fitting curve in terms of the moments calculated from the given statistics, and it will be convenient now to make our approach from the other end.

Take the general equation of the fitting curve, express its constants in terms of its moments, and substitute for the latter the values determined from the statistics, since the basis of the fitting is the equalization of the moments of the observational curve and of the theoretical curve. This will enable us to determine κ, the criterion for fixing the type of curve suitable to the given distribution. When the type has been fixed it is, as a rule, not a very difficult matter to express the constants of the particular type again in terms of the observational moments.

Now the general differential equation of the fitting curve was

$$\frac{dy}{dx}=\frac{y(x+b)}{px^2+qx+r};$$

hence

$$\int(px^2+qx+r)dy=\int y(x+b)dx,$$

where the integration is to traverse the complete curve.

Therefore, multiplying both sides by x^n,

$$\int(px^{n+2}+qx^{n+1}+rx^n)dy=\int(yx^{n+1}+byx^n)dx;$$

or, if we integrate the left-hand side by parts

$$\left[(px^{n+2}+qx^{n+1}+rx^n)y\right]-\int y(\overline{n+2}px^{n+1}+\overline{n+1}qx^n+nrx^{n-1})dx$$
$$=\int(yx^{n+1}+byx^n)dx.$$

But the expression in square brackets vanishes at both limits if we suppose y to be zero at each end of the curve, so that the equation reduces to

$$(1+p\overline{n+2})\int yx^{n+1}dx+(b+q\overline{n+1})\int yx^ndx+rn\int yx^{n-1}dx=0, \quad . \quad . \quad . \quad (9)$$

Now if deviations are measured from the mean of the distribution, we have

$$\int yxdx = N\mu_1 = 0,\ \int yx^2dx = N\mu_2,\ \int yx^3dx = N\mu_3,\ \text{etc.},$$

and therefore, putting $n=3$ in the above relation,

$$(1+5p)N\mu_4 + (b+4q)N\mu_3 + 3rN\mu_2 = 0\ ;$$

put $n=2$, $$(1+4p)N\mu_3 + (b+3q)N\mu_2 = 0\ ;$$

put $n=1$, $$(1+3p)N\mu_2 + rN = 0\ ;$$

put $n=0$, $$(b+q)N = 0.$$

Thus $b=-q$, and, on substitution in the other three equations, we get

$$\begin{aligned} 5\mu_4 p + 3\mu_3 q + 3\mu_2 r + \mu_4 &= 0,\\ 4\mu_3 p + 2\mu_2 q \qquad\quad + \mu_3 &= 0,\\ 3\mu_2 p \qquad + \qquad r + \mu_2 &= 0, \end{aligned}$$

three simple linear equations to find p, q, r, the solution of which leads to

$$\begin{aligned} p &= -(2\mu_2\mu_4 - 3\mu^2{}_3 - 6\mu^3{}_2)/(10\mu_2\mu_4 - 18\mu^3{}_2 - 12\mu^2{}_3),\\ q &= -b = -\mu_3(\mu_4 + 3\mu^2{}_2)/(10\mu_2\mu_4 - 18\mu^3{}_2 - 12\mu^2{}_3),\\ r &= -\mu_2(4\mu_2\mu_4 - 3\mu^2{}_3)/(10\mu_2\mu_4 - 18\mu^3{}_2 - 12\mu^2{}_3). \end{aligned}$$

We have thus expressed p, q, r, and b, the constants of the fitting curve in terms of the moments of the observed distribution, but the results may be rendered more concise by writing

$$\beta_1 = \mu^2{}_3/\mu^3{}_2,\ \beta_2 = \mu_4/\mu^2{}_2, \quad . \quad . \quad . \quad (10)$$

whence

$$p = -(2\beta_2 - 3\beta_1 - 6)/2(5\beta_2 - 6\beta_1 - 9), \quad . \quad . \quad . \quad . \quad (11)$$

$$q = -b = -\sqrt{(\mu_2\beta_1)}\,.\,(\beta_2+3)/2(5\beta_2 - 6\beta_1 - 9), \quad . \quad . \quad (12)$$

$$r = -\mu_2(4\beta_2 - 3\beta_1)/2(5\beta_2 - 6\beta_1 - 9) \quad . \quad . \quad . \quad . \quad (13)$$

And κ, the criterion for fixing the type of curve suitable to the statistics given, is immediately deduced from

$$\begin{aligned} \kappa &= q^2/4pr\\ &= \beta_1(\beta_2+3)^2/4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6) \quad . \quad . \quad . \quad (14) \end{aligned}$$

Also, since $\frac{dy}{dx}$ vanishes when $x=-b$, this fixes the mode relative to the origin. But the origin is now at the mean, so that

$$\text{mode} - \text{mean} = -b = -\sqrt{(\mu_2\beta_1)}\,.\,(\beta_2+3)/2(5\beta_2 - 6\beta_1 - 9) \quad (15)$$

And

$$\begin{aligned} \text{skewness} &= (\text{mean} - \text{mode})/\text{S.D.}\\ &= b/\sqrt{(\mu_2)}\\ &= \sqrt{\beta_1}(\beta_2+3)/2(5\beta_2 - 6\beta_1 - 9) \quad . \quad . \quad . \quad (16) \end{aligned}$$

CHAPTER XVII

APPLICATIONS OF CURVE FITTING

WE are in a position now to test the application of these principles to given frequency distributions and we shall start by trying to find a curve to fit the record of marks obtained by 514 candidates in a certain examination (see p. 25).

Example (1).—This example is chosen because it turns out, when we come to evaluate κ, that it is well fitted by the *normal curve, Type VII*, which is one of the simplest and at the same time the most important of all the types discussed. Before we start the numerical part of the work it will be well to express the constants y_0 and σ of this curve in terms of the moments of the distribution.

The equation of the normal curve is

$$y = y_0 e^{-\frac{x^2}{2\sigma^2}}.$$

If N be the total frequency, we have by equation (4) *bis*, p. 202,

$$N = \int_{-\infty}^{+\infty} y dx$$

$$= y_0 \int_{-\infty}^{+\infty} e^{-x^2/2\sigma^2} dx.$$

Put $x^2/2\sigma^2 = \xi^2$, so that $\frac{dx}{d\xi} = \sigma\sqrt{2}$ and when $x = \infty$, $\xi = \infty$ also.

Thus
$$N = y_0 \sigma \sqrt{2} \int_{-\infty}^{+\infty} e^{-\xi^2} d\xi$$

$$= y_0 \sigma \sqrt{2} \sqrt{\pi} \text{ (see Appendix, Note 8)}$$

$$= \sqrt{(2\pi)}\sigma y_0 \quad . \quad . \quad . \quad . \quad (1)$$

Again $$\mu_2 = \int_{-\infty}^{+\infty} yx^2dx \Big/ \int_{-\infty}^{+\infty} ydx$$

$$= \frac{y_0}{N}\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2}x^2dx$$

$$= \frac{2y_0}{N} \cdot \sigma\sqrt{2} \cdot 2\sigma^2 \int_0^{\infty} e^{-\xi^2}\xi^2 d\xi$$

$$= -\frac{2\sqrt{2} \cdot \sigma^3 y_0}{N}\int_0^{\infty} \xi d(e^{-\xi^2})$$

$$= -\frac{2\sqrt{2} \cdot \sigma^3 y_0}{N}\left[\left(\xi e^{-\xi^2}\right)_0^{\infty} - \int_0^{\infty} e^{-\xi^2}d\xi\right]$$

$$= \frac{2\sqrt{2} \cdot \sigma^3 y_0}{N} \cdot \frac{\sqrt{\pi}}{2},$$

since $\left(\xi e^{-\xi^2}\right)_0^{\infty}$ vanishes at both limits.

Therefore, $\mu_2 = \sqrt{2} \cdot \sigma y_0 \sqrt{\pi} \cdot \sigma^2/N = \sigma^2$, by (1).

In fact, σ **is simply the S.D. of the distribution.**

And $\mathbf{y_0 = N/\sqrt{(2\pi)} \cdot \sigma}$.

TABLE (41). DISTRIBUTION OF MARKS OBTAINED BY 514 CANDIDATES IN A CERTAIN EXAMINATION.

Mean No. of Marks.	Deviation from 33.	Frequency of Candidates.	First Moment.	Second Moment.	Third Moment.	Fourth Moment.
	(x)	(f)	(fx)	(fx^2)	(fx^3)	(fx^4)
3	−6	5	− 30	180	−1080	6480
8	−5	9	− 45	225	−1125	5625
13	−4	28	−112	448	−1792	7168
18	−3	49	−147	441	−1323	3969
23	−2	58	−116	232	− 464	928
28	−1	82	− 82	82	− 82	82
33	..	87	..	..	..	..
38	+1	79	+ 79	79	+ 79	79
43	+2	50	+100	200	+ 400	800
48	+3	37	+111	333	+ 999	2997
53	+4	21	+ 84	336	+1344	5376
58	+5	6	+ 30	150	+ 750	3750
63	+6	3	+ 18	108	+ 648	3888
—	—	514	−110	2814	−1646	41,142

The first 4 moments referred to 33 as origin and with the class interval, 5 marks, as unit of deviation, are

$$-110/514,\ 2814/514,\ -1646/514,\ 41142/514.$$

The arithmetic mean of the distribution

$$\begin{aligned}&=33+5\bar{x}\\&=33+5(-\tfrac{110}{514})\\&=33-5(0{\cdot}214008)\\&=31{\cdot}92996.\end{aligned}$$

The second, third, and fourth moments referred to the mean as origin, and retaining five marks as unit of deviation, are given (see Appendix, Note 5) by

$$\begin{aligned}\nu_2&=2814/514-\bar{x}^2=5{\cdot}42891\\\nu_3&=-1646/514-3\bar{x}\nu_2-\bar{x}^3=0{\cdot}29296\\\nu_4&=41142/514-4\bar{x}\nu_3-6\bar{x}^2\nu_2-\bar{x}^4=78{\cdot}79964.\end{aligned}$$

After making Sheppard's adjustments

$$\mu_2=\nu_2-\tfrac{1}{12},\ \mu_3=\nu_3,\ \mu_4=\nu_4-\tfrac{1}{2}\nu_2+\tfrac{7}{240},$$

these become

$$\mu_2=5{\cdot}34558,\ \mu_3=0{\cdot}29296,\ \mu_4=76{\cdot}11436.$$

Thus $\beta_1=\mu^2{}_3/\mu^3{}_2=0{\cdot}00056,\ \beta_2=\mu_4/\mu^2{}_2=2{\cdot}66365.$

Hence

$$\begin{aligned}\kappa&=\beta_1(\beta_2+3)^2/4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)\\&=(0{\cdot}00056)(5{\cdot}66365)^2/4(10{\cdot}65292)(-0{\cdot}67438)\\&=-0{\cdot}00063.\end{aligned}$$

Since κ and β_1 are small and β_2 does not differ greatly from 3, making p and q small, we may fit a normal curve to this distribution.

The appropriate normal curve is

$$y=y_0e^{-x^2/2\sigma^2},$$

where $\sigma^2=\mu_2=5{\cdot}34558$ (5 marks as unit),

$$y_0=\mathrm{N}/\sqrt{(2\pi\mu_2)}=514/\sqrt{2\pi}(5{\cdot}34558)^{\frac{1}{2}}=88{\cdot}6903.$$

Hence the required curve has for its equation, writing results to three significant figures,

$$y=88{\cdot}7e^{-x^2/2(5{\cdot}35)}.$$

Now the mean of the distribution is at 31·92996, where the central ordinate of the normal curve is erected, and the distance of any x, say x_{33}, from this point

$$\begin{aligned}&=(33-31{\cdot}92996)/5 \text{ (expressed with 5 marks as unit)}\\&=0{\cdot}214008.\end{aligned}$$

Any other x may be found in the same way and y can then be deduced from the equation of the curve by taking logs, thus

$$\log_{10} y = \log_{10} 88{\cdot}6903 - \frac{x^2}{2(5{\cdot}34558)} \log_{10} e$$

$$= 1{\cdot}9478762 - (0{\cdot}0406218)x^2.$$

This enables us to calculate the ordinates of the normal curve and thence we could evaluate the areas by successive applications of a suitable quadrature formula.

We can, however, get the areas direct by using a table of the probability integral, such as that due to Dr. W. F. Sheppard (see pp. 284, 285). In that case the corresponding abscissæ have first to be expressed in terms of the standard deviation as unit, *e.g.*

$$x_{40{\cdot}5} = 40{\cdot}5 - 31{\cdot}92996 = 8{\cdot}57004,$$

and $$\sigma = 5\sqrt{(5{\cdot}34558)} = 11{\cdot}56025,$$

where the factor 5 is introduced because 5 marks was the unit in the calculation of μ_2 (a process equivalent in effect to that previously adopted).

Thus $x_{40{\cdot}5}/\sigma = 0{\cdot}741336$
$= \xi$, say.

The area of the normal curve up to the abscissa x/σ or ξ

$$= \int_{-\infty}^{x} y\,dx$$

$$= \int_{-\infty}^{x} y_0 e^{-x^2/2\sigma^2} dx$$

$$= y_0 \int_{-\infty}^{\xi} e^{-\xi^2/2} \sigma\, d\xi$$

$$= y_0\sigma \,.\, \sqrt{2\pi} \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi$$

$$= \mathrm{N} \int_{-\infty}^{\xi} z\, d\xi$$

$$= \mathrm{N} \,.\, \tfrac{1}{2}(1+\alpha),$$

where $\frac{\alpha}{2}$ represents the area of the curve $z = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}$ between 0 and ξ.

Sheppard's Tables give the values of $\frac{1}{2}(1+\alpha)$ for different values of ξ, and when

$$\xi=0{\cdot}74,\ \tfrac{1}{2}(1+\alpha)=0{\cdot}7703500$$
$$\xi=0{\cdot}75,\ \tfrac{1}{2}(1+\alpha)=0{\cdot}7733726.$$

Therefore, by interpolation, when

$$\xi=0{\cdot}741336,\ \tfrac{1}{2}(1+\alpha)=0{\cdot}7707538.$$

Thus the frequency of candidates with marks lying between 0 and 40·5

$$=514(0{\cdot}7707538)=396{\cdot}17.$$

Similarly the frequency of candidates with marks lying between 0 and 45·5=452·20.

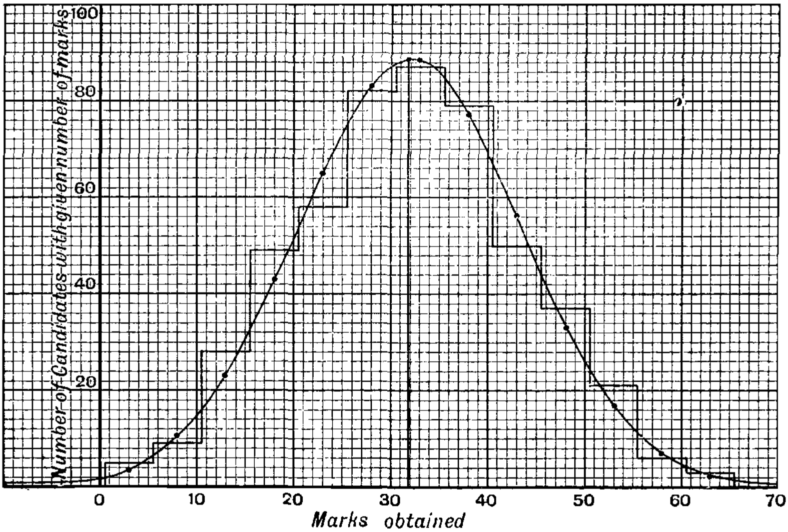


Fig. (35).

Hence the normal frequency for the group with 43 as mean number of marks=56·0, and the same method gives the area for any other group.

The histogram of the observations and the curve plotted from the ordinates are shown together in fig. (35).

In Table (42) are set out the calculated normal frequency (col. (4)) for each group alongside the corresponding observed frequency (col. (2)), and the differences between the two are shown in col. (5). We want to know whether the fit is a good one.

TABLE (42). COMPARISON OF OBSERVED AND NORMAL FREQUENCIES IN EXAMINATION EXAMPLE.

(1) Mean No. of Marks.	(2) Observed Frequency.	(3) Normal Frequency. Ordinates.	(4) Normal Frequency. Areas.	(5) Deviation.	(6) Sq. of Deviation.	(7) Ratio of No. in Col. (6) to No. in Col. (4).
3	5	3·9	5·7	+0·7	0·49	0·09
8	9	10·4	10·7	+1·7	2·89	0·27
13	28	23·2	23·5	−4·5	20·25	0·86
18	49	42·9	43·1	−5·9	34·81	0·81
23	58	65·8	65·6	+7·6	57·76	0·88
28	82	83·7	83·1	+1·1	1·21	0·01
33	87	88·3	87·6	+0·6	0·36	0·00
38	79	77·3	76·8	−2·2	4·84	0·06
43	50	56·1	56·0	+6·0	36·00	0·64
48	37	33·7	34·0	−3·0	9·00	0·26
53	21	16·8	17·1	−3·9	15·21	0·89
58	6	7·0	7·2	+1·2	1·44	0·20
63	3	2·4	3·5	+0·5	0·25	0·07
..	514	511·5	513·9	..	184·51	$\chi^2=5{\cdot}04$

Now with this object we might square each difference as in col. (6), sum the squares, and find the mean square deviation by dividing by the total frequency; this, after extracting the square root, would give what might be called the root-mean-square error, regarding the theoretical values as the true ones. In the above example it

$$=\sqrt{(184{\cdot}51/514)}=0{\cdot}599.$$

But this form of result, while it may be useful in some cases, *e.g.* in comparing two distributions of the same kind to some theoretical series, is open to objection; for one thing it treats all the differences as if they were of equal importance in absolute magnitude, but a difference of 2, say, in a normal frequency of 10 is clearly more serious than a like difference in a frequency of 60. The objection, however, goes deeper than that; even when the root-mean-square deviation is found we are at a loss to estimate its precise relationship to the quality of fit, as there seems to be no definite connection between one distribution and another of a different kind: there is no standard case, so to speak, to which we can always appeal, where the fit is agreed to be good and supplying therefore a suitable root-mean-square deviation for comparison.

This leads us to the question: What constitutes goodness of fit? Suppose by some means we have selected a theoretical or empirical formula to describe a certain frequency distribution in a given population; if the frequency values observed do not differ from the theoretical frequencies by more than the deviations we might expect owing to random sampling, then clearly the fit may be regarded as a good one. And we have a measure of the fit if we can find the proportion of random samples, of the same size as the given distribution, showing greater deviations from the distribution given by theory than those which are actually observed.

Now Professor Karl Pearson has shown how this proportion can be calculated [*Phil. Mag.*, vol. l., pp. 157-175 (1900)]; he finds the probability that a random sample should give a frequency distribution differing from that which theory proposes *by as much as or by more than* the distribution actually observed. This probability, **P**, is a function of χ, where

$$\chi^2 = \Sigma[(y-y')^2/y],$$

y and y' representing the theoretical and observed frequencies for any particular group and the summation is to include all groups. It will be noted that this expression gives each difference $(y-y')$ its appropriate importance by relating it to the frequency y of its own group.

A table in *Biometrika* (vol. i., pp. 155 *et seq.*) gives the values of P corresponding to different values of χ^2 (including all integral values from 1 to 30) and to values of n', the total number of frequency groups, from 3 to 30 (see also p. 285). The mathematics involved in finding P is difficult, and the reader who wishes to enter into it must consult the original memoir, but the utility of the function has been proved by experience and it is readily applied in a particular case.

In the above example χ^2 is found from col. (7): it equals 5·04, and from the table of values of P, when $n'=13$, we have

$$P=0{\cdot}957979 \text{ when } \chi^2=5,$$

and

$$P=0{\cdot}916082 \text{ when } \chi^2=6.$$

Therefore, by proportional interpolation, when $\chi^2=5{\cdot}04$, $P=0{\cdot}956303$. Thus, supposing our data to follow the normal curve, in 956 random samples out of 1000 we should expect to get a worse-fitting distribution than that given by the sample actually observed. We may therefore conclude without hesitation that the normal curve provides an excellent fit in this particular instance.

We pass on now to fresh distributions to illustrate some of the other types of frequency curve.

Example (2) deals with the percentage of trade union members unemployed at the end of each month for the years 1898 to 1912 [data from the *Sixteenth Abstract of Labour Statistics of the United Kingdom*, Cd. 7131]. Table (43) shows the distribution of the 180 records according to the percentage unemployed.

The deviations are measured from the centre of the group (3·9–5·2) as origin, and the class interval (1·3 per cent.) is taken as unit of deviation as usual.

The first four moments are :—

$$-29/180(=\bar{x}),\ 425/180,\ 397/180,\ 3053/180\,;$$

i.e. $$-0{\cdot}1611111,\ 2{\cdot}3611111,\ 2{\cdot}2055556,\ 16{\cdot}9611111.$$

TABLE (43). DISTRIBUTION OF UNEMPLOYED PERCENTAGES OF TRADE UNION MEMBERS

Percentage Unemployed.	Deviation.	Frequency.	First Moment.	Second Moment.	Third Moment.	Fourth Moment.
0—	−3	0	0	0	0	0
1·3—	−2	33	−66	132	−264	528
2·6—	−1	57	−57	57	− 57	57
3·9—	..	41	..	..	..	..
5·2—	+1	24	+24	24	+ 24	24
6·5—	+2	10	+20	40	+ 80	160
7·8—	+3	11	+33	99	+297	891
9·1—	+4	3	+12	48	+192	768
10·4—	+5	1	+ 5	25	+125	625
..	..	180	−29	425	+397	3053

Referred to the mean,

$$4{\cdot}55+1{\cdot}3\bar{x}=4{\cdot}3405556,$$

the second, third, and fourth moments are (see Appendix, Note 5),

$$\nu_2=2{\cdot}3611111-\bar{x}^2=2{\cdot}3351543,$$
$$\nu_3=2{\cdot}2055556-3\bar{x}\nu_2-\bar{x}^3=3{\cdot}338395,$$
$$\nu_4=16{\cdot}9611111-4\bar{x}\nu_3-6\bar{x}^2\nu_2-\bar{x}^4=18{\cdot}74817.$$

Owing to the very doubtful contact at the beginning of the curve Sheppard's adjustments were not made in this case, but the rough moments as calculated above were used.

Thus $\beta_1 = \nu^2_3/\nu^3_2 = 0{\cdot}875242$

$\beta_2 = \nu_4/\nu^2_2 = 3{\cdot}43817$

and $\kappa = \beta_1(\beta_2+3)^2/4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6) = -0{\cdot}466.$

Since κ is negative the fitting curve should be of *Type I.*, the equation of which is

$$\mathbf{y} = \mathbf{y}_0\left(1+\frac{\mathbf{x}}{\mathbf{a}_1}\right)^{\mathbf{m}_1}\left(1-\frac{\mathbf{x}}{\mathbf{a}_2}\right)^{\mathbf{m}_2}$$

where $m_1/a_1 = m_2/a_2$, and $(a_1+a_2) = b$, say.

It is therefore necessary before going further to determine y_0, a_1, a_2, b, m_1 and m_2 in terms of ν_2, ν_3, ν_4, or β_1 and β_2, the constants of the distribution.

The value of y_0 is found to be most conveniently expressed as a *Gamma* function which is defined, with the usual notation, thus :—

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx,$$

whence it follows that $\Gamma(k+1) = k\Gamma(k)$. [See Appendix, Note 9, also p. 285.]

Also, if

$$B(m, n) = \int_0^1 x^{m-1}\ (1-x)^{n-1}dx$$

it may be easily shown that

$B(m, n) = \Gamma(m)\Gamma(n)/\Gamma(m+n)$. [See Appendix, Note 9.]

The general method of procedure in determining the constan for all the different types is :—

1. Express the fact that the area of the curve is a measure the total frequency of the distribution—this enables us find y_0.
2. Find the nth moment of the curve with regard to some fi origin—giving n particular values, 1, 2, 3, 4, this lead the determination of μ_2, μ_3, μ_4, β_1, β_2 in terms of the stants of the curve, and thence to formulæ for calcula the constants.

Once found, the same formulæ may be used, of course, i cases of the same type : we have only to replace letters b numbers for which they stand.

Applying this method to the Type I. curve, we have

$$N = \int_{-a_1}^{+a_2} ydx$$

$$= \frac{y_0}{a_1^{m_1}a_2^{m_2}}\int_{-a_1}^{+a_2}(a_1+x)^{m_1}(a_2-x)^{m_2}dx$$

Put $(a_1+x)=(a_1+a_2)z$, so that $(a_2-x)=(a_1+a_2)(1-z)$ and $\frac{dx}{dz}=(a_1+a_2)=b$; therefore

$$N=\frac{y_0 b(a_1+a_2)^{m_1+m_2}}{a_1^{m_1}a_2^{m_2}}\int_0^1 z^{m_1}(1-z)^{m_2}dz \quad . \quad . \quad . \quad (2)$$

$$=\frac{y_0 b(m_1+m_2)^{m_1+m_2}}{m_1^{m_1}m_2^{m_2}}B(m_1+1,\ m_2+1).$$

Hence $\quad \mathbf{y_0=\frac{N}{b}\cdot\frac{m_1^{m_1}m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}}\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}}.$

Again, $N\mu'_n=\int_{-a_1}^{+a_2}y(a_1+x)^n dx$

is the nth moment of the distribution referred to $(-a_1, 0)$, the point where the curve starts from the axis on the left-hand side, as origin.

Therefore, as above,

$$N\mu'_n=\frac{y_0}{a_1^{m_1}a_2^{m_2}}\int_{-a_1}^{+a_2}(a_1+x)^{m_1+n}(a_2-x)^{m_2}dx$$

$$=\frac{by_0(a_1+a_2)^{m_1+n+m_2}}{a_1^{m_1}a_2^{m_2}}\int_0^1 z^{m_1+n}(1-z)^{m_2}dz$$

$$=b^nN\int_0^1 z^{m_1+n}(1-z)^{m_2}dz\Big/\int_0^1 z^{m_1}(1-z)^{m_2}dz, \text{ by (2)}.$$

Hence,

$$\mu'_n=b^n\Gamma(m_1+n+1)\Gamma(m_1+m_2+2)/\Gamma(m_1+1)\Gamma(m_1+m_2+n+2)$$

$$=b^n(m_1+n)(m_1+n-1)\ .\ .\ .\ (m_1+1)/(m_1+m_2+n+1)(m_1+m_2+n$$
$$\ .\ .\ .\ (m_1+m_2+2),$$

by repeated application of the relation $\Gamma(k+1)=k\Gamma(k)$.

Putting $n=1, 2, 3, 4$ in succession, we have

$$\mu'_1=b(m_1+1)/(m_1+m_2+2),$$
$$\mu'_2=b^2(m_1+2)(m_1+1)/(m_1+m_2+3)(m_1+m_2+2),$$
$$\mu'_3=b^3(m_1+3)(m_1+2)(m_1+1)/(m_1+m_2+4)(m_1+m_2+3)(m_1+m_2+$$
$$\mu'_4=b^4(m_1+4)(m_1+3)(m_1+2)(m_1+1)/(m_1+m_2+5)(m_1+m_2+4)$$
$$(m_1+m_2+3)(m_1+m_2+2).$$

These relations are rendered more concise if we write

$$m_1+1=m'_1,\ m_2+1=m'_2,\ m_1+m_2+2=r\ ;$$

thus

$$\mu'_1=bm'_1/r$$
$$\mu'_2=b^2m'_1(m'_1+1)/r(r+1)$$
$$\mu'_3=b^3m'_1(m'_1+1)(m'_1+2)/r(r+1)(r+2)$$
$$\mu'_4=b^4m'_1(m'_1+1)(m'_1+2)(m'_1+3)/r(r+1)(r+2)(r+3$$

To get the corresponding moments referred to the **mean as** origin we have the relations :—

$$\mu_1=0, \qquad \mu_3=\mu'_3-3\mu'_1\mu_2-\mu'^3_1,$$
$$\mu_2=\mu'_2-\mu'^2_1, \qquad \mu_4=\mu'_4-4\mu'_1\mu_3-6\mu'^2_1\mu_2-\mu'^4_1,$$

which, after some straightforward reduction, give

$$\mu_2=b^2m'_1m'_2/r^2(r+1)$$
$$\mu_3=2b^3m'_1m'_2(m'_2-m'_1)/r^3(r+1)(r+2)$$
$$\mu_4=3b^4m'_1m'_2[m'_1m'_2(r-6)+2r^2]/r^4(r+1)(r+2)(r+3).$$

Thus $\beta_1=\mu^2_3/\mu^3_2=\dfrac{4b^6m'^2_1m'^2_2(m'_2-m'_1)^2}{r^6(r+1)^2(r+2)^2}\Big/\dfrac{b^6m'^3_1m'^3_2}{r^6(r+1)^3}$

$$=4(m'_2-m'_1)^2(r+1)/m'_1m'_2(r+2)^2$$
$$=4(r^2-4m'_1m'_2)(r+1)/m'_1m'_2(r+2)^2.$$

Therefore, $\dfrac{r^2}{m'_1m'_2}=\dfrac{\beta_1(r+2)^2}{4(r+1)}+4$ . . . (3)

Again, $\beta_2=\mu_4/\mu^2_2=\dfrac{3b^4m'_1m'_2[m'_1m'_2(r-6)+2r^2]}{r^4(r+1)(r+2)(r+3)}\Big/\dfrac{b^4m'^2_1m'^2_2}{r^4(r+1)^2}$

$$=\frac{3[m'_1m'_2(r-6)+2r^2]}{m'_1m'_2}\cdot\frac{(r+1)}{(r+2)(r+3)}.$$

Therefore, $\dfrac{2r^2}{m'_1m'_2}=-r+6+\dfrac{\beta_2(r+2)(r+3)}{3(r+1)}$ . . . (4)

Combining (3) and (4), $\dfrac{2\beta_1(r+2)^2}{4(r+1)}+8=6-r+\beta_2\dfrac{(r+2)(r+3)}{3(r+1)}$,

whence $\mathbf{r=6(\beta_2-\beta_1-1)/(3\beta_1-2\beta_2+6)}$ . . . (5)

Again, since $\mu_2=b^2m'_1m'_2/r^2(r+1)$,

therefore $b^2=\mu_2(r+1)\,.\,[\beta_1(r+2)^2+16(r+1)]/4(r+1)$, by (3),

i.e. $\mathbf{b=\tfrac{1}{2}\sqrt{\mu_2}\sqrt{[\beta_1(r+2)^2+16(r+1)]}}$ . . (6)

And $m'_1m'_2=4r^2(r+1)/[\beta_1(r+2)^2+16(r+1)]$,

while $m'_1+m'_2=r$; hence m'_1 and m'_2 are roots of

$$m^2-rm+\frac{4r^2(r+1)}{\beta_1(r+2)^2+16(r+1)}=0,$$

the solution of which quadratic is $\dfrac{r}{2}\pm\tfrac{1}{2}\sqrt{\left[r^2-\dfrac{16r^2(r+1)}{\beta_1(r+2)^2+16(r+1)}\right]}$;

therefore, m_2 **and** m_1* are respectively equal to

$$\frac{1}{2}\left[r-2\pm\frac{r(r+2)\sqrt{\beta_1}}{\sqrt{\{\beta_1(r+2)^2+16(r+1)\}}}\right] \quad . \quad . \quad . \quad (7)$$

and a_1 and a_2 follow from

$$\frac{a_1}{m_1}=\frac{a_2}{m_2}=\frac{b}{m_1+m_2} \quad . \quad . \quad . \quad . \quad (8)$$

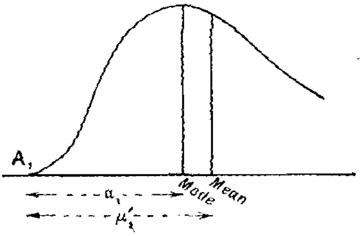
Applying these formulæ to the 'unemployed' example, we find

$r=5{\cdot}36048.$	$m_1=0{\cdot}169185.$	$m_2=3{\cdot}191295.$
$b=9{\cdot}33236.$	$a_1=0{\cdot}469842.$	$a_2=8{\cdot}86252.$

Also $y_0=58{\cdot}1282$, and the equation of the curve is therefore

$$y=58{\cdot}1\left(1+\frac{x}{0{\cdot}470}\right)^{0{\cdot}169}\left(1-\frac{x}{8{\cdot}86}\right)^{3{\cdot}19}.$$

The position of the origin, which is at the mode, is given by



$$\begin{aligned}(\text{mean}-\text{mode})&=\mu'_1-a_1\\ &=\frac{bm'_1}{r}-\frac{bm_1}{m_1+m_2}\\ &=b\left(\frac{m'_1}{r}-\frac{m'_1-1}{r-2}\right)\\ &=b\,\frac{m'_2-m'_1}{r(r-2)}\\ &=\frac{1}{2}\frac{\mu_3}{\mu_2}\frac{r+2}{r-2}; \quad . \quad . \quad . \quad . \quad (9)\end{aligned}$$

thus, $\text{mode}=4{\cdot}3405556-\frac{1}{2}\cdot\frac{\nu_3}{\nu_2}\cdot\frac{r+2}{r-2},$

in this particular case,

$$=2{\cdot}3052009.$$

[* When μ_3 is positive m_2 goes with the positive root of the quadratic, and *vice versa*.]

This enables us to write down any x, and thence y by substituting for x in the equation of the curve, which, by taking logs, may be written

$$\log y = \log y_0 + m_1 \log\left(1+\frac{x}{a_1}\right) + m_2 \log\left(1-\frac{x}{a_2}\right);$$

e.g. for the x of the group (2·6—3·9), bearing in mind that **1·3** is the unit of measurement for x, we have

$$x_{3\cdot 25} = (3\cdot 25 - 2\cdot 3052009)/1\cdot 3 = 0\cdot 9447991/1\cdot 3.$$

Hence $\left(1+\frac{x_{3\cdot 25}}{a_1}\right) = 2\cdot 546835$; $\left(1-\frac{x_{3\cdot 25}}{a_2}\right) = 0\cdot 9179953$;

$m_1 \log\left(1+\frac{x_{3\cdot 25}}{a_1}\right) = 0\cdot 0686892$; $m_2 \log\left(1-\frac{x_{3\cdot 25}}{a_2}\right) = -0\cdot 118587$;

so that $\log y = 1\cdot 714489$,

and $y_{3\cdot 25} = 51\cdot 82$.

Similarly the ordinates at the centre points of the other groups may be calculated, but it must be remembered that the resulting values are only a first approximation to the observed frequencies, and a better series is obtained if, by using some good quadrature formula, we calculate the *areas* for the successive groups between the curve, the bounding ordinates, and the axis of x. Indeed in the case of the group (1·3—2·6) it is essential to do this, because (1) the rise of the curve is so very abrupt as to render the determination of the single ordinate at the centre quite inadequate for an accurate measure of the frequency in that group, and (2) a portion of the group falls outside the range of the curve which only starts at 1·6944063 (*i.e.* mode$-1\cdot 3a_1$), and this has to be allowed for in finding the frequency as represented by the area between the curve and axis.

The base of the required area, range (1·6944063 to 2·6), was therefore divided into eight equal parts and the ordinates at the points of division were determined. The area was then found by using Simpson's well-known formula :—

$$\text{Area} = \tfrac{1}{3}h[(y_0+y_{2p}) + 2(y_2+y_4+\ldots+y_{2p-2}) + 4(y_1+y_3+\ldots+y_{2p-1})],$$

where h denotes the length of one of the equal parts into which the base is divided and $2p$ is their number ; in our case $p=4$ and $h=\frac{1}{8}$, the class interval being the unit, and the result is to be reduced in the ratio

$$0\cdot 9055937 : 1\cdot 3$$

in order to allow for the smaller range of this group ; we thus get as the area for the group

$$\frac{0\cdot9055937}{1\cdot3}\times\frac{1}{24}[(y_0+y_8)+2(y_2+y_4+y_6)+4(y_1+y_3+y_5+y_7)]=37\cdot39.$$

The observed and calculated frequencies for the whole series are compared in Table (44), the remaining areas in col. (4) being calculated by the simpler but somewhat less accurate form of Simpson's formula, when only three ordinates are used, namely,

$$\int_{-1}^{+1} ydx=\tfrac{1}{3}(y_{-1}+4y_0+y_1).$$

TABLE (44). COMPARISON OF OBSERVED AND THEORETICAL FREQUENCIES OF UNEMPLOYED PERCENTAGES

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Percentage Unemployed.	Observed Frequency.	Theoretical Frequency.		*Deviation.*	Square of Deviation.	Ratio of No. in *Col. (6) to* No. in Col. (4).
		Ordinates.	Areas.			
1·3—	33	55·3*	37·4	+4·4	19·36	0·52
2·6—	57	51·8	51·6	−5·4	29·16	0·57
3·9—	41	37·8	37·8	−3·2	10·24	0·27
5·2—	24	24·9	25·0	+1·0	1·00	0·04
6·5—	10	14·8	14·9	+4·9	24·01	1·61
7·8—	11	7·7	7·8	−3·2	10·24	1·31
9·1—	3	3·3	3·4	+0·4	0·16	0·05
10·4—	1	1·0	1·2	+0·2	0·04	0·03
..	180	..	179·1	..	..	$\chi^2=4\cdot40$

To test the goodness of fit we have $n'=8$, $\chi^2=4\cdot40$, whence, by means of the P table, P=0·731852. Thus, roughly, we may say that three out of every four random samples of 180 records would give a worse fit with the proposed curve than is given by the actual distribution observed, so that the fit may be regarded as quite a reasonably good one. This conclusion is also supported by an examination of the curve which has been drawn, fig. (36), with the histogram of the given statistics.

Example (3).—The data for this example concerning infectious diseases will be found in Table (16), p. 62 (or, see p. 224) ; the reader should work out the moments for himself and verify the following results :—

[* The ordinate in this case cannot be accepted as an approximation to the frequency given by the curve.]

The first four moments referred to 7 as origin are

$$0{\cdot}282158,\quad 4{\cdot}86307,\quad 17{\cdot}4855,\quad 129{\cdot}394.$$

Referred to the mean, 7·564316, the three latter become

$$\nu_2=4{\cdot}78346,\quad \nu_3=13{\cdot}4140,\quad \nu_4=111{\cdot}964.$$

If we do not assume high contact at the terminals, and certainly at the lower end it is doubtful, we deduce from the above values of the moments that

$$\beta_1=1{\cdot}64396,\quad \beta_2=4{\cdot}89321,\quad \kappa=-1{\cdot}53.$$

Thus the fitting curve is of *Type I.* and its constants, when calculated, are

$$r=11{\cdot}7819.\quad m_1=0{\cdot}31171.\quad m_2=9{\cdot}47020.$$
$$a_1=0{\cdot}79216.\quad a_2=24{\cdot}0671.\quad y_0=60{\cdot}363.$$

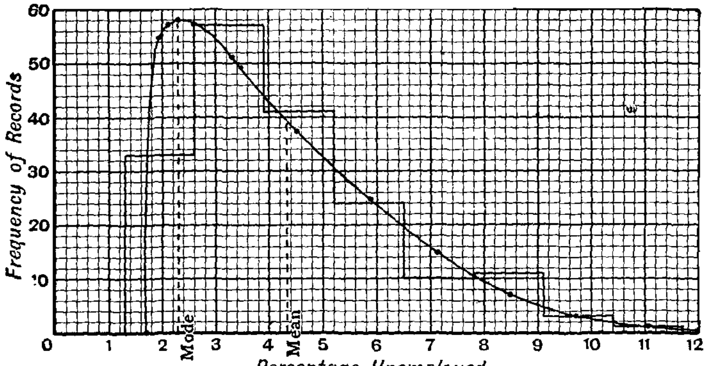


Fig. (36).

The equation of the curve is therefore, retaining three significant figures throughout

$$y=60{\cdot}4\left(1+\frac{x}{0{\cdot}792}\right)^{0{\cdot}312}\left(1-\frac{x}{24{\cdot}1}\right)^{9{\cdot}47}.$$

The curve starts at 2·02904 (so that the first group of observations lies wholly outside its range) and ends at 51·7475. It is drawn, together with the corresponding histogram, in fig. (37).

Supposing, just for the sake of comparison, *we assume high contact* at the terminals and attempt to fit the given distribution with a Type III. curve, to which Type I. is closely related.

We then have, after making Sheppard's adjustments,

$$\mu_2=4{\cdot}70013,\quad \mu_3=13{\cdot}4140,\quad \mu_4=109{\cdot}601,$$

whence $\beta_1=1{\cdot}73295,\quad \beta_2=4{\cdot}96129,\quad \kappa=-1{\cdot}47.$

It will be noted that the theoretically correct type to take here again is Type I., but this was discarded because, when attempted,

it led to a curve starting at a point corresponding to a disease rate of 3·385, so that the central ordinates of each of the first two observed groups lay outside the curve altogether.

Type III. curve is of the form

$$\mathbf{y}=\mathbf{y}_0\mathbf{e}^{-\gamma\mathbf{x}}\left(1+\frac{\mathbf{x}}{\mathbf{a}}\right)^{\gamma\mathbf{a}}.$$

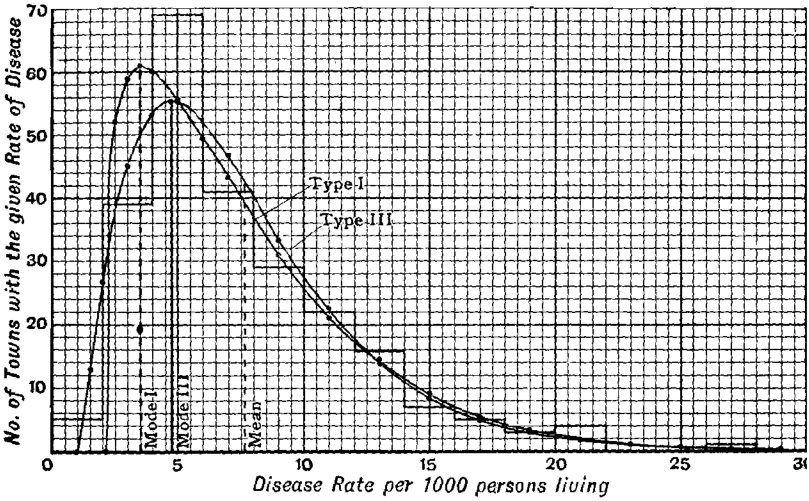


Fig. (37).

To express the constants in terms of the moments, noting that the curve starts from $x=-a$ on one side and goes off to infinity on the other, we have

$$\begin{aligned}
N&=\int_{-a}^{\infty}y\,dx\\
&=y_0\int_{-a}^{\infty}e^{-\gamma x}\left(1+\frac{x}{a}\right)^{\gamma a}dx\\
&=\frac{y_0}{a^p}\int_{-a}^{\infty}e^{-\gamma x}(a+x)^p dx \text{ (where } \gamma\mathbf{a}=\mathbf{p})\\
&=\frac{y_0}{a^p\gamma^p}\int_{-a}^{\infty}e^{-\gamma x}(\gamma a+\gamma x)^p dx\\
&=\frac{y_0}{p^p}e^{\gamma a}\int_{-a}^{\infty}e^{-(\gamma a+\gamma x)}(\gamma a+\gamma x)^p dx\\
&=\frac{y_0e^p}{\gamma p^p}\int_{0}^{\infty}e^{-z}z^p dz \text{ (where } \gamma a+\gamma x=z)\\
&=\frac{ay_0e^p}{p^{p+1}}\Gamma(p+1).
\end{aligned}$$

Therefore, $$\mathbf{y}_0=\mathbf{N}\mathbf{p}^{\mathbf{p}+1}/\mathbf{a}\mathbf{e}^{\mathbf{p}}\Gamma(\mathbf{p}+1) \quad . \quad . \quad . \quad (10).$$

Again, the nth moment of the distribution referred to $(-a, 0)$ as origin is

$$N\mu'_n=\int_{-a}^{\infty} y(a+x)^n dx$$

$$=\frac{y_0}{a^p}\int_{-a}^{\infty} e^{-\gamma x}(a+x)^{p+n}dx$$

$$=\frac{y_0}{a^p}\cdot\frac{e^p}{\gamma^{p+n}}\int_{-a}^{\infty} e^{-(\gamma a+\gamma x)}(\gamma a+\gamma x)^{p+n}dx$$

$$=\frac{y_0}{a^p}\ \frac{e^p}{\gamma^{p+n+1}}\int_0^{\infty} e^{-z}z^{p+n}dz.$$

Therefore, by (10),

$$\mu'_n=\frac{p^{p+1}}{ae^p\Gamma(p+1)}\cdot\frac{e^p}{a^p\gamma^{p+n+1}}\cdot\Gamma(p+n+1)$$

$$=\Gamma(p+n+1)/\gamma^n\Gamma(p+1).$$

Hence,

$$\mu'_1=\Gamma(p+2)/\gamma\Gamma(p+1)=(p+1)/\gamma$$

$$\mu'_2=\Gamma(p+3)/\gamma^2\Gamma(p+1)=(p+2)(p+1)/\gamma^2$$

$$\mu'_3=\Gamma(p+4)/\gamma^3\Gamma(p+1)=(p+3)(p+2)(p+1)/\gamma^3.$$

Transferring to the mean as origin we have for the moments, since

$$\bar{x}=\mu'_1=(p+1)/\gamma$$

$$\mu_2=\mu'_2-\bar{x}^2=(p+1)/\gamma^2$$

$$\mu_3=\mu'_3-3\bar{x}\mu_2-\bar{x}^3=2(p+1)/\gamma^3.$$

Hence, combining these last two equations,

$$\gamma=2\mu_2/\mu_3, \qquad p=(4\mu_2^3/\mu_3^2)-1 \quad . \quad . \quad . \quad (11)$$

In our particular case these equations give

$$\gamma=0{\cdot}700780, \quad p=1{\cdot}30820, \quad a=1{\cdot}86678,$$

and, therefore, by (10),

$$y_0=55{\cdot}3323.$$

Hence the curve is

$$y=55{\cdot}3e^{-0{\cdot}701x}\left(1+\frac{x}{1{\cdot}87}\right)^{1{\cdot}31}.$$

The equation of the curve, on taking logs, gives

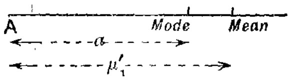
$$\log y=\log y_0-\gamma\log_{10}e\,.\,x+p\log\left(1+\frac{x}{a}\right)$$

$$=1{\cdot}742979-0{\cdot}304345x+1{\cdot}30820\log(1+x/1{\cdot}86678).$$

Before we can go on to calculate the ordinates of the curve we need to know where the origin lies, and since it coincides with the mode it may be found from

$$\begin{aligned} \text{mean}-\text{mode} &= \mu'_1 - a \\ &= (p+1)\gamma - p/\gamma \\ &= \frac{1}{\gamma} = \frac{\mu_3}{2\mu_2} \quad . \quad . \quad . \quad . \quad . \quad (12) \end{aligned}$$

Thus, $\text{mode} = 7{\cdot}564316 - 2{\cdot}853960 = 4{\cdot}71036.$



Suppose now we wish to calculate the ordinate **corresponding to** the x of the centre point of group (6—8), we have

$$\begin{aligned} x_7 &= \tfrac{1}{2}(7-4{\cdot}71036) \\ &= 1{\cdot}14482, \end{aligned}$$

bearing in mind that the unit is a rate of 2 per 1000.

Hence, substituting this value in the equation for log y,

$$\begin{aligned} \log y_7 &= 1{\cdot}666278 \\ y_7 &= 46{\cdot}374, \end{aligned}$$

and similarly any other y may be found.

The curve starts at

$$\text{mode} - a = 4{\cdot}71036 - 2(1{\cdot}86678) = 0{\cdot}97680,$$

so that the range of the first group as determined from the curve is (0·9768—2), and not (0—2) as in the observations.

The ordinates and afterwards the areas, calculated by a method somewhat similar to that indicated in Example (2), were determined for each separate group of observations, and the results for both Type I. and Type III. curves are compared in Table (45).

Type III. curve is drawn on the same diagram, fig. (37), as Type I. curve and the observation histogram, and the result lends emphasis to an important point, namely, the necessity for replacing ordinates by areas to obtain the frequency proper to any group.

In order to get a measure of the goodness of fit in each case, the function P was calculated, but in the Type I. comparison the first group had to be omitted to avoid the infinite term which would have resulted in χ^2_1, owing to this group falling right outside the curve, that is to say, the test had to be confined to towns in which

the observed case rate was not less than 2. The values found for P were :—

Type I.—P=0·34307,
Type III.—P=0·46298,.

so that in every 100 samples containing 241 observations each, we should get, roughly, 34 deviating from the Type I. curve and 46 deviating from the Type III. curve, at least as widely as the given distribution. In neither case can the fit be regarded as a very good one, but the failure is only marked in one or two groups, such as that of maximum frequency, where there may be other than random causes to account for it ; *e.g.* where isolation is inefficient the disease is likely to spread, one case infects another : in other words, the events are not independent.

TABLE (45). COMPARISON OF OBSERVED DISTRIBUTION OF INFECTIOUS DISEASE RATES, NOTIFIED IN 241 LARGE TOWNS OF ENGLAND AND WALES, WITH THEORETICAL DISTRIBUTION.

(1)	(2)	(3)	(4)	(5)	(6)
Case Rate.	Observed Frequency.	Theoretical Frequency.		$(f_1-f)^2/f_1$.	$(f_3-f)^2/f_3$.
		Type I.	Type III.		
	(f)	(f_1)	(f_3)		
0—	5	..	6·6	..	0·39
2—	39	52·6	43·7	3·52	0·51
4—	69	55·4	54·3	3·34	3·98
6—	41	43·2	46·2	0·11	0·59
8—	29	31·2	33·6	0·15	0·63
10—	22	21·5	22·4	0·01	0·01
12—	16	14·2	14·1	0·23	0·26
14—	7	9·1	8·6	0·48	0·30
16—	5	5·6	5·1	0·06	0·00
18—	3	3·3	2·9	0·03	0·00
20—	4	1·9	1·7	2·32	3·11
22—	0	1·0	0·9	1·00	0·90
24—	0	0·5	0·5	0·50	0·50
26—	1	0·3	0·3	1·63	1·63
..	241	239·8	240·9	$\chi^2_1=13·38$	$\chi^2_3=12·81$

Example (4) refers to the wages of certain women tailors previously recorded in Table (11), p. 41. The data as given in the original suffered a disadvantage common to such statistics : at

either end the grouping differed from that in the centre, two or three classes being lumped together owing to the smallness of frequency in each. The figures ran thus :—Under 5s., 19 ; 5s. and under 6s., 180 ; 6s. and under 7s., 384 ; . . . ; 23s. and under 24s., 64 ; 24s. and under 25s., 54 ; 25s. and under 30s., 122 ; 30s. and over, 36. They were recast in the form shown in Table (46), suggested by an examination of the histogram, in order to make the fitting simpler.

The first four moments calculated from this adapted table and referred to 12s. as origin are :—

$$\nu'_1=0{\cdot}556718,\ \nu'_2=5{\cdot}056373,\ \nu'_3=16{\cdot}70163,\ \nu'_4=123{\cdot}7691.$$

When referred to the mean, 13·113436, the last three become

$$\nu_2=4{\cdot}746438,\quad \nu_3=8{\cdot}60179,\quad \nu_4=95{\cdot}6914\ ;$$

or, after making Sheppard's adjustments,

$$\mu_2=4{\cdot}663105,\quad \mu_3=8{\cdot}60179,\quad \mu_4=93{\cdot}3474\ ;$$

therefore, $\beta_1=0{\cdot}729713,\quad \beta_2=4{\cdot}29291,\quad \kappa=1{\cdot}63.$

The curve is thus of *Type VI.*,

$$\mathbf{y}=\mathbf{y}_0(\mathbf{x}-\mathbf{a})^{q_2}/\mathbf{x}^{q_1}.$$

To calculate the constants, the nth moment about the origin is given by

$$\begin{aligned}
\mathrm{N}\mu'_n&=\int_a^\infty yx^n dx\\
&=y_0\int_a^\infty (x-a)^{q_2}x^{n-q_1}dx\\
&=y_0\int_1^0 a^{q_2}\frac{(1-z)^{q_2}}{z^{q_2}}\cdot\frac{a^{n-q_1}}{z^{n-q_1}}a\left(-\frac{1}{z^2}\right)dz\ \left(\text{where } x=\frac{a}{z}\right)\\
&=\frac{y_0}{a^{q_1-q_2-n-1}}\int_0^1 z^{q_1-q_2-n-2}(1-z)^{q_2}dz\\
&=\frac{y_0}{a^{q_1-q_2-n-1}}\mathrm{B}(q_1-q_2-n-1,\ q_2+1).
\end{aligned}$$

Thus, putting $n=0$,

$$\mathrm{N}=\frac{y_0}{a^{q_1-q_2-1}}\mathrm{B}(q_1-q_2-1,\ q_2+1)\quad .\quad .\quad .\quad (13)$$

and $\qquad \mu'_n=a^n\Gamma(q_1-q_2-1-n)\Gamma(q_1)/\Gamma(q_1-n)\Gamma(q_1-q_2-1)$;

therefore, $\quad \mu'_1=a\Gamma(q_1-q_2-2)\Gamma(q_1)/\Gamma(q_1-1)\Gamma(q_1-q_2-1)$
$\qquad\qquad =a(q_1-1)/(q_1-q_2-2).$

Also $\mu'_n/\mu'_{n-1}=a\Gamma(q_1-q_2-1-n)\Gamma(q_1-n+1)/\Gamma(q_1-n)\Gamma(q_1-q_2-n)$
$\qquad\qquad =a(q_1-n)/(q_1-q_2-n-1).$

Hence $\mu'_2=a^2(q_1-1)(q_1-2)/(q_1-q_2-2)(q_1-q_2-3)$

$$\mu'_3=a^3(q_1-1)(q_1-2)(q_1-3)/(q_1-q_2-2)(q_1-q_2-3)(q_1-q_2-4)$$
$$\mu'_4=a^4(q_1-1)(q_1-2)(q_1-3)(q_1-4)/(q_1-q_2-2)(q_1-q_2-3)(q_1-q_2-4)(q_1-q_2-5).$$

But these relations are precisely the same as those of Type I. with a in place of b, $-q_1$ in place of m_1, and q_2 in place of m_2, so that $(1+q_2)$, $(1-q_1)$* are the roots of

$$q^2-rq+4r^2(r+1)/[\beta_1(r+2)^2+16(r+1)]=0 \quad . \quad . \quad (14)$$

where $$r=6(\beta_2-\beta_1-1)/(6+3\beta_1-2\beta_2) \quad . \quad . \quad . \quad . \quad (15)$$

Also $$y_0=Na^{q_1-q_2-1}\Gamma(q_1)/\Gamma(q_1-q_2-1)\Gamma(q_2+1), \text{ by (13)} \quad . \quad (16)$$

and a is given by

$$\mu_2=a^2(1-q_1)(1+q_2)/r^2(r+1), \quad . \quad . \quad . \quad . \quad (17)$$

μ_2 being the second moment of the given distribution referred to its mean as origin.

The distance of the mean from the origin is

$$\mu'_1=a(q_1-1)/(q_1-q_2-2),$$

and this fixes the origin, for the mean is known directly from the statistics.

To get the mode, use the equation of the curve, putting $\frac{dy}{dx}=0$ and we have

$$\text{origin}=\text{mode}-aq_1/(q_1-q_2).$$

Combining this with

$$\text{origin}=\text{mean}-a(q_1-1)/(q_1-q_2-2)$$

we have

$$\text{mean}-\text{mode}=a(q_1+q_2)/(q_1-q_2)(q_1-q_2-2) \quad . \quad . \quad (18)$$

Applying these formulæ to the case of the women tailors,

$$r=-38{\cdot}7698, \quad q_1=51{\cdot}5269, \quad q_2=10{\cdot}7571, \quad a=21{\cdot}11018,$$

and the equation of the curve is

$$y=y_0(x-21{\cdot}1)^{10{\cdot}8}/x^{51{\cdot}5},$$

where $\log y_0=68{\cdot}8254$.

Also the origin is at $-41{\cdot}9104$, the mode at $11{\cdot}4498$, and the maximum theoretical frequency is 2299.

[* When μ_3 is positive $(1+q_2)$ goes with the positive root of the quadratic, and *vice versa.*]

TABLE (46). DISTRIBUTION OF WAGES OF CERTAIN WOMEN TAILORS, ACTUAL AND THEORETICAL.

Wages.	Frequency. Actual.	Frequency. Theoretical.	Wages.	Frequency. Actual.	Frequency. Theoretical.
1s.—	5	1	19s.—	523	503
3s.—	14	52	21s.—	262	278
5s.—	564	452	23s.—	118	147
7s.—	1243	1332	25s.—	64	75
9s.—	2045	2096	27s.—	43	38
11s.—	2339	2255	29s.—	27	19
13s.—	1815	1898	31s.—	15	9
15s.—	1432	1353	33s.—	9	5
17s.—	854	859	..	..	..
..	..	..	..	11,372	11,372

The theoretical and actual frequencies are compared in Table (46) and the curve is drawn with the histogram in fig. (38).

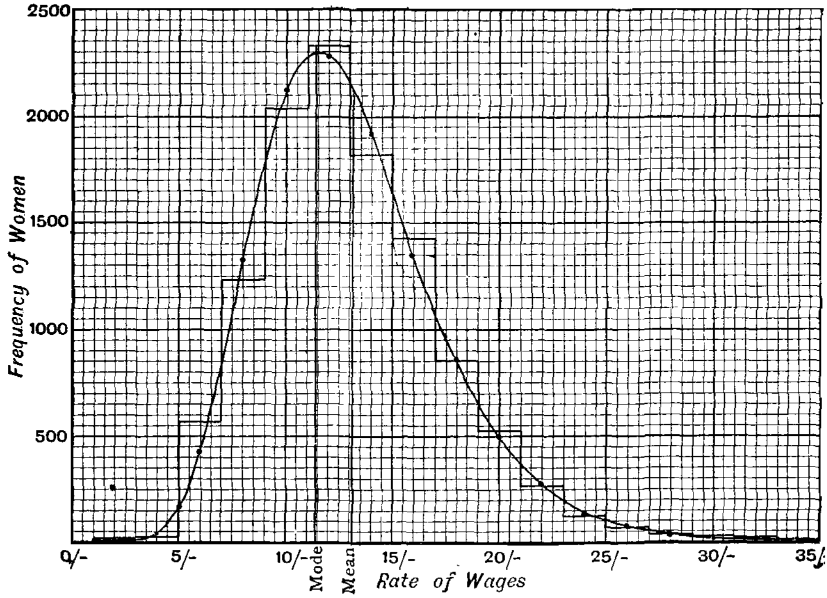


FIG. (38).

Example (5) discusses the distribution of frequencies of specimens of Anemone nemorosa with different numbers of sepals, recorded by G. U. Yule (*Biometrika*, vol. i., p. 307).

The first four moments referred to 6 as origin are

$$\nu'_1=0{\cdot}508,\quad \nu'_2=1{\cdot}012,\quad \nu'_3=2{\cdot}476,\quad \nu'_4=9{\cdot}124.$$

Referred to the mean, 6·508, the last three become

$$\nu_2=0{\cdot}7539360,\quad \nu_3=1{\cdot}195905,\quad \nu_4=5{\cdot}459941.$$

The contact, at one extremity certainly, being doubtful, Sheppard's adjustments were not made in this case. Hence,

$$\beta_1=3{\cdot}337259,\quad \beta_2=9{\cdot}605476,\quad \kappa=1{\cdot}46.$$

Since κ does not differ greatly from unity an attempt was made to fit the observations with a *Type V.* curve, namely,

$$\mathbf{y}=\mathbf{y}_0\mathbf{x}^{-\mathbf{p}}\mathbf{e}^{-\gamma/\mathbf{x}}.$$

The nth moment about the origin is given by

$$N\mu'_n=\int_0^\infty yx^n dx$$

(since, p and γ being positive, y vanishes at $x=0$ and at $x=\infty$)

$$=y_0\gamma^{n-p+1}\int_0^\infty z^{p-n-2}e^{-z}dz \text{ (where } z=\gamma/x)$$

$$=y_0\gamma^{n-p+1}\Gamma(p-n-1).$$

Thus $\quad N=y_0\gamma^{-p+1}\Gamma(p-1).$

And $\mu'_n/\mu'_{n-1}=\gamma/(p-n-1).$

Hence $\quad \mu'_1=\gamma/(p-2)$

$$\mu'_2=\gamma^2/(p-2)(p-3)$$

$$\mu'_3=\gamma^3/(p-2)(p-3)(p-4).$$

Referred to the mean as origin, the last two moments become

$$\mu_2=\gamma^2/(p-2)^2(p-3),$$

$$\mu_3=4\gamma^3/(p-2)^3(p-3)(p-4),$$

whence

$$\beta_1=\mu^2_3/\mu^3_2=16(p-3)/(p-4)^2=[16(p-4)+16]/(p-4)^2;$$

this gives a quadratic for $(p-4)$, one solution of which is

$$\mathbf{p}-4=[8+4\sqrt{(4+\beta_1)}]/\beta_1, \quad . \quad . \quad . \quad (19)$$

the positive root being taken in order to get a real γ.

Thus $\quad \gamma^*=(\mathbf{p}-2)\sqrt{[(\mathbf{p}-3)\mu_2]} \quad . \quad . \quad . \quad (20)$

and $\quad \mathbf{y}_0=\mathbf{N}\gamma^{\mathbf{p}-1}/\Gamma(\mathbf{p}-1) \quad . \quad . \quad . \quad . \quad (21)$

Since $\quad \mu'_1=\gamma/(p-2)$, the position of the origin is given by

$$\text{Origin}=\text{Mean}-\gamma/(\mathbf{p}-2) \quad . \quad . \quad . \quad (22)$$

Also the distance of the mode from the origin is $\gamma/\mathbf{p}$, so that all the constants of the curve are readily determined.

[* The sign of γ is taken to be the same as that of μ_3.]

In our particular case, we get

$$p=9\cdot643840, \quad \gamma=17\cdot10758,$$

and the curve is

$$y=y_0x^{-9\cdot64}e^{-17\cdot1/x},$$

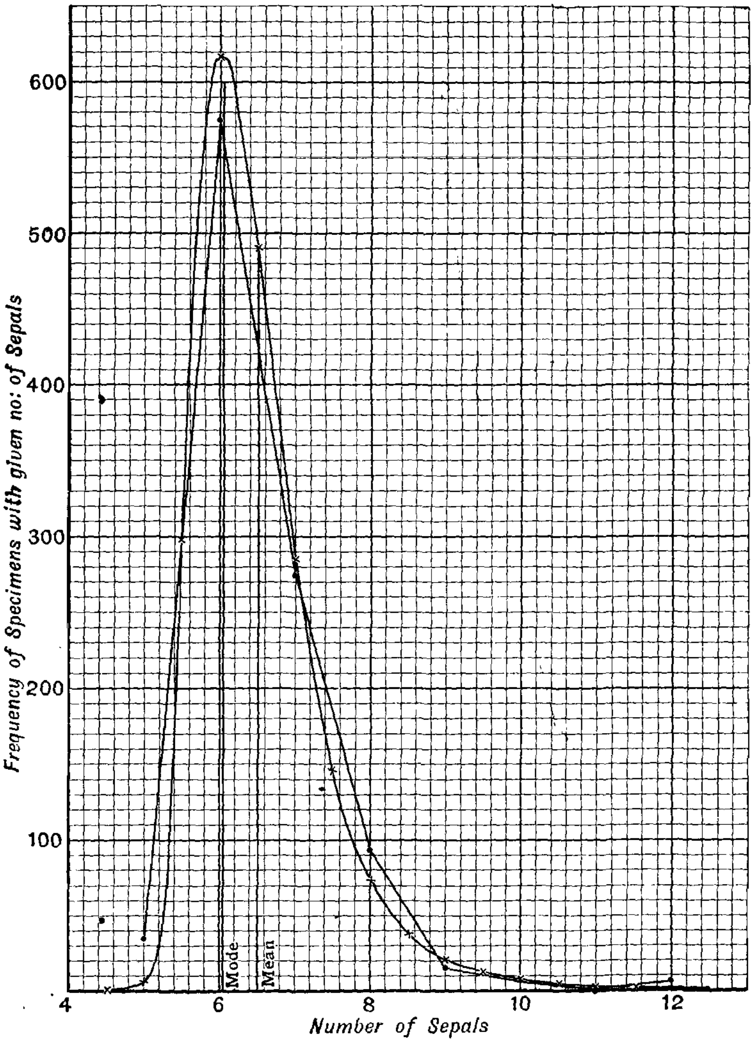


FIG. (39).

where $\log y_0=9\cdot38179$. The origin is at 4·27 and the mode at 6·04. The greatest frequency is 620 approximately, and the frequency distribution, calculating areas for the several groups as if they ranged between (4·5—5·5), (5·5—6·5), etc., is shown alongside the observed

distribution in Table (47). The curve is plotted in fig. (39) from the ordinates which were calculated at the centre and extremities of each group so as to enable Simpson's simple quadrature formula to be used to get the areas.

TABLE (47). DISTRIBUTION OF SEPALS OF ANEMONE NEMOROSA, OBSERVED AND CALCULATED.

No. of Sepals.	Frequency.		No. of Sepals.	Frequency.	
	Observed.	Calculated.		Observed.	Calculated.
5	34	51	9	14	22
6	576	544	10	4	6
7	276	296	11	..	2
8	92	81	12	4	1
..	..	..	..	1000	1003

[Examples have been given above of five out of the seven different types of frequency curve that have been enumerated. For further examples of all the types and a complete account of the method reference should be made to *Professor Pearson's* memoirs, especially the following :—

Roy. Soc. Phil. Trans., vol. 186A, pp. 343-414 (1895), *On Skew Variation in Homogeneous Material;* and a *Supplementary Memoir* in vol. 197A, pp. 443-459 (1901).

Biometrika, vol. i., pp. 265 *et seq.*, *On the Systematic Fitting of Curves to Observations and Measurements*, continued in vol. ii., pp. 1-23. Also vol. iv., pp. 169-212, which discusses various historical hypotheses made to generalize the Gaussian Law, the basis of the symmetrical normal curve.

A large number of highly interesting practical illustrations of Pearsonian curve fitting occur throughout the pages of *Biometrika*, while W. P. Elderton's *Frequency Curves and Correlation* contains an admirably concise treatment of the theory, with applications to meet more particularly the actuarial point of view.

It should be stated that rival curves and methods have been proposed as suitable for fitting certain types of frequency distribution, some of which have scarcely received the attention and the trial they deserve. Among the most interesting are those developed by *Professor Edgeworth* ; for some account of his voluminous work upon the subject the reader may refer to several memoirs in the *Journal of the Royal Statistical Society*, beginning December 1898 (the *Method of Translation*), among which the following are important as giving more recent results of his researches :—

Vol. lxix. (1906), *The Generalized Law of Error or Law of Great Numbers.*

Vol. lxxvii. (1914), *On the Use of Analytical Geometry to Represent Certain Kinds of Statistics.*

Vol. lxxix. (1916), *On the Mathematical Representations of Statistical Data ;* continued in vol. lxxx. (1917).

Two memoirs may be cited as of particular interest—those of May 1917 and March 1918—because they reply to criticism and draw a comparison from their author's point of view between his curves and those of Professor Pearson.]

CHAPTER XVIII

THE NORMAL CURVE OF ERROR

LET us return for a moment to the general statement on p. 143, that 'whenever we have n similar but independent events happening in which the probability of success for each is p, the different resulting possibilities as to success are given by the successive terms in $(s+f)^n$, namely,

$$s^n+ns^{n-1}f^1+\frac{n(n-1)}{1 \cdot 2}s^{n-2}f^2+ \ldots +f^n,$$

and their correspondent probabilities by the successive terms in $(p+q)^n$, namely,

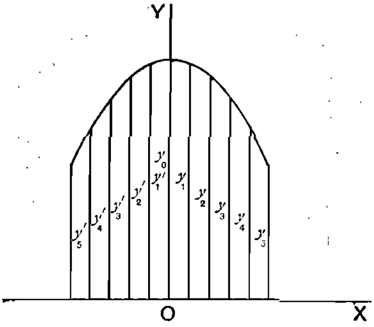
$$p^n+np^{n-1}q^1+\frac{n(n-1)}{1 \cdot 2}p^{n-2}q^2+ \ldots +q^n.\text{'}$$

When we come to try and apply this theory directly to cases other than those of random sampling in artificial experiments with coins, dice, etc., we are faced at once with difficulties because of the limiting character of the assumption on which the theory rests, namely, that *all the events are to be similar and independent.* The similarity demanded is of the same radical type as that existing when we throw the same die or spin the same coin twice running, and the test for it is that p, the chance of success, is to be the same for every individual event. The independence is to be such that no single event and no combination of events is to have any influence upon any of the rest.

Now for most classes of events it is impossible to assign any *a priori* value to p at all, still less can we be sure that p does not change from one event to the next. For example, the chance of death for soldiers in war-time varies from regiment to regiment according to where they happen to be located; for the same regiment it varies from battalion to battalion according to whether they are in the trenches or behind the lines; and from individual to individual according to innumerable little accidents of time, place,

and condition. Also, where the shells burst thickest, p increases for any soldier there, but it increases also for his neighbour. Thus the events in such a case are not similar, neither are they independent.

Moreover, as it stands, the theory cannot be applied to any distribution in which the character observed is capable of *continuous* variation. This difficulty, however, has been overcome, as we have seen, by replacing the histogram representative of the binomial by a continuous curve which at the same time serves to describe the discontinuous series to a high degree of accuracy.



To illustrate how close this description can be, even when n is comparatively small, we will fit with its appropriate normal curve the symmetrical binomial polygon formed by joining up the summits of the ordinates representing successive terms of the series

$$2^{10}(\tfrac{1}{2}+\tfrac{1}{2})^{10},$$

erected at unit distance apart.

The total area bounded by the polygon, the extreme ordinates, and the axis of x is practically

$$=(y_0+y_1+y_2+\ \ldots\ +y'_1+y'_2+\ \ldots)\times(1)$$
$$=\text{sum of the given ordinates}$$
$$=2^{10}(\tfrac{1}{2}+\tfrac{1}{2})^{10}$$
$$=1024.$$

The equation of the normal curve is

$$y=\mathrm{Y}_0 e^{-x^2/2\sigma^2},$$

where $\sigma^2=npq=11\times\tfrac{1}{2}\times\tfrac{1}{2}=2{\cdot}75,$

and $\mathrm{Y}_0=\mathrm{N}/\sqrt{2\pi}\cdot\sigma=1024/\sqrt{(5{\cdot}5\pi)}.$

Hence, taking logs, we have

$$\log y=\log \mathrm{Y}_0-\frac{x^2}{2\sigma^2}\log_{10} e$$

$$=2{\cdot}3915437-x^2(0{\cdot}0789626).$$

It is easy from this equation to calculate the normal curve ordinates corresponding to $x=0, 1, 2, 3, 4, 5,$ and the results, compared with the polygon ordinates, are as follows :—

x	Ordinate of Polygon.	Normal Curve Ordinate.
0	252	246·3
±1	210	205·4
±2	120	119·0
±3	45	48·0
±4	10	13·4
±5	1	2·6

Now although the circumstances in which the series

$$(\tfrac{1}{2})^n+n(\tfrac{1}{2})^{n-1}(\tfrac{1}{2})+\frac{n(n-1)}{1\,.\,2}(\tfrac{1}{2})^{n-2}(\tfrac{1}{2})^2+\ \ldots$$

may be taken to represent the frequency distribution resulting from a particular kind of experiment were so stringently defined, there is no reason why the normal curve itself to which the theory led should be subjected to precisely the same limitations. After all, the real and only justification for choosing one curve rather than another to fit any given observations is that it does succeed in fitting them better. But when the further question is asked *why* the normal curve should succeed in describing some results so well, we must not be tempted by analogy to rush to the conclusion that the causes at work are necessarily independent, and equal, and so on. In short, the theoretical justification and the empirical use of the normal curve are two quite different matters.

Experience shows that the normal curve suffices to fit certain types of distribution, besides those which arise in tossing coins and in similar experiments, with remarkable accuracy; among these may be noted :—

1. *Certain biological statistics ;* for instance, the proportions of male to female births taken over a series of years for a large community such as the population of a country; also the proportions of different types of plants and animals resulting from cross-fertilization.

2. *Certain anthropometrical, particularly craniometrical and allied statistics*, such as the height, weight, lengths of various bones, skull measurements, etc., of a large group of persons, and the agreement is the closer if the group be reasonably homogeneous, *i.e.* composed of individuals of the same nationality and sex between the same narrow age limits, etc.; also measurements of a similar character in animals and plants.

3. *Errors of observation in experimental work*; for example,

several measurements of the same quantity—length, weight, speed, temperature, or whatever it be—will contain errors of this kind which are equally liable to be above or below the true value.

4. *The marks of shots upon a given target*, assuming that the shots are equally liable to err in any given direction. This is an interesting case of the normal law in two dimensions, for the north and south line and the east and west line through the centre of the target may both be regarded as axes of normal curves of error.*

5. *Certain sociological statistics of a comparatively stationary character;* for example, rates of birth, marriage, or death at neighbouring times or like places; also the wages (and possibly the output if it could be satisfactorily measured) of large numbers of workers engaged in the same occupation under the same general conditions.

6. *Any statistics or quantities that are individually compounded of large number of elements*, mostly independent of one another, which themselves vary between limits not very widely divergent, and none of which exert a preponderating influence upon their resultant statistic. The latter may be simply the sum of its elements, or, more generally, it may be any function of the elements which, to the first degree of approximation, can be expressed in linear form.

Now it would be a difficult matter in most of these cases to satisfy ourselves as to the fulfilment or non-fulfilment of conditions like those on which the binomial distribution rests. It is not easy indeed to visualize them perfectly, except in artificial experiments where they are largely under control. If anything, the chances seem almost hopelessly against their fulfilment in ordinary life, so closely must we hedge round our sample to keep out unequal influences. For example, to use a frequently quoted illustration, if p measures the chance of death for an individual, the death rate varies, as we know, considerably from place to place according to the age and sex constitution of the population; it is influenced by differences in class, and occupation, and manner of life; it is altered from time to time, violently by the ravages of war or disease, more gradually by improvement in general sanitation, housing conditions, etc. We should only expect to get the binomial distribution (and consequently the normal law if it depended upon the

[* Sir John Herschel published in the *Edinburgh Review* (1850) an *a priori* proof of the normal law from a consideration of this problem. Taking $\phi(x^2)$ as the expression of the law for one dimension and $\phi(x^2+y^2)$ for two dimensions, the independence of errors in perpendicular directions leads to the functional equation $\phi(x^2+y^2)=\phi(x^2)\times\phi(y^2)$, the solution of which is of the form $\phi(x^2)=\frac{h}{\sqrt{\pi}}e^{-h^2x^2}$. It should be added that the assumptions underlying the proof are not entirely above criticism.]

same postulates) exactly verified if we were dealing with the same stationary population existing under the same stable conditions over a long period of time ; moreover, since p is to be identical for each individual event in the ideal case, it would be further necessary that every family and every individual in our population should also remain in the same stationary and stable state. This is manifestly impossible, especially after the industrial revolution which the advent of machine power created.

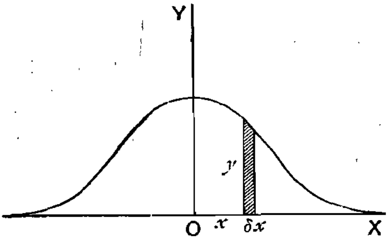
These considerations suggest the interesting question whether the various types of statistics we have enumerated, as being approximately subject to the normal law, could not, if we knew more about them, really all be included under heading number (6), representing a further development from the binomial theory and an enlargement of the field in which it holds good.

In an earlier chapter, when we were discussing the connection between marriage rate and prices, we showed how it was possible by a method of averaging to differentiate between long-time and short-time effects. The more transient fluctuations, only superficial in character, were removed and the real nature of any permanent change in the figures was revealed. In much the same way, when we have a group of statistics which do not perhaps fit a normal curve of error at all closely, it may be possible by random averaging to get rid of some of the fluctuations which cause the badness of fit and to obtain a new group of statistics which more nearly obey the normal law. Averaging, that is to say, tends to smooth away the rough outstanding abnormalities ; and we shall presently show that if two variables, X_1, X_2, which are independent, obey the normal law, any linear function of the variables $(w_1X_1+w_2X_2)$, obeys the same law. This may throw some light on Class (6) where each statistic represents a compound, that is, in a broad sense, a kind of an average of a large number of elements which partially neutralize one another's influence, or rub the corners off one another, so to speak, since no single element is, by hypothesis, to exert an overwhelming influence upon the compound itself.

But although the normal curve does serve to describe a considerable number of frequency distributions within reasonable limits, there are many more cases in which it fails : for example, the greater part of those bearing on economic matters ; also statistics relating to the incidence of disease and degree of fertility are, as a rule, very markedly skew. Hence arose the necessity for an extension from the symmetrical normal to some kind of skew variation curves to fit such distributions.

The normal curve, however, has an importance of its own to which we must now draw special attention. *It is the foundation of the theory of errors* and provides us with an invaluable method of estimating the importance of one error in comparison with another, or of determining the probability that an error shall lie between stated limits. Upon it we depend for several most important approximations which are in constant use.

The term ' error ' is used here in the sense that if we take the mean of a number of observations, the deviation of any one of them from the mean may be termed its error. When such deviations can be satisfactorily fitted, that is, within the limits of random sampling, by means of a normal curve, they are said to be subject to the normal law of error. This law is expressed, as we have seen, by the equation



$$y=\frac{N}{\sqrt{2\pi}\cdot\sigma}e^{-x^2/2\sigma^2},$$

where $y\,.\,\delta x$ measures the frequency with which an observed organ or character deviates from the mean by an amount lying between x and $(x+\delta x)$ in a large population, *i.e.* $y\,.\,\delta x$ registers the frequency of an error of size x to $(x+\delta x)$, and N and σ are constants dependent upon the particular application of the law.

The probability curve or normal curve of error. As a guide to the drawing of the above curve it may be worth while plotting

$$y=e^{-x^2}.$$

This is readily done by writing the equation in the form

$$-x^2=\log_e y.$$

Giving now to y the values 0, 0·1, 0·2, etc., we can find values of $\log_e y$ as shown in Table (48), and, by means of a square root table, x is then determined.

TABLE (48). CORRESPONDING VALUES OF x AND y TO PLOT $y=e^{-x^2}$.

y	$\log_e y$	x	y	$\log_e y$	x
0	$-\infty$	$\pm\infty$	0·6	−0·5108	±0·71
0·1	−2·3026	±1·52	0·7	−0·3567	±0·60
0·2	−1·6095	±1·27	0·8	−0·2232	±0·47
0·3	−1·2040	±1·10	0·9	−0·1054	±0·32
0·4	−0·9163	±0·96	1·0	0	0
0·5	−0·6932	±0·83	..	..	..

This enables us to plot the graph as shown in fig. (40). Since $\log_e 1=0$, and the logarithm of any number greater than 1 is positive and thus cannot be equal to $-x^2$, it follows that y cannot be greater than 1. Moreover y cannot be less than 0, for the logarithm of a negative quantity is meaningless, but, as y approaches 0, x approaches ∞.

Also the curve is symmetrical about OY because for any possible value of y there are two values of x, equal and opposite.

Returning now to the curve

$$y=\frac{N}{\sqrt{2\pi}\,.\,\sigma}\,e^{-x^2/2\sigma^2},$$

it must be of the same general shape as $y=e^{-x^2}$ because the two only differ in their constants. It is clearly symmetrical, for

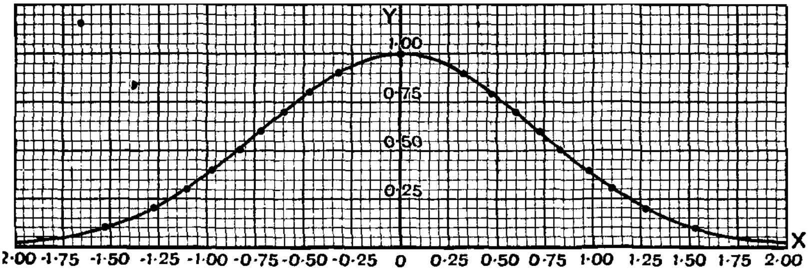


FIG. (40). The graph of $y=e^{-x^2}$.

instance, about the axis of y, because, in this case also, to any value of y there are two values of x equal and opposite. Moreover it tails off to the right and left from OY, the axis of x being an asymptote, for as x tends to $\pm\infty$, y tends to zero as before.

When $\qquad x=0, \quad y=N/\sqrt{2\pi}\,.\,\sigma,$

giving the point B, fig. (41), where the curve cuts the axis of y. This is evidently the highest point on the curve, for

$$\frac{dy}{dx}=-\frac{Nx}{\sqrt{2\pi}\,.\,\sigma^3}\,e^{-x^2/2\sigma^2},$$

and this vanishes when $x=0$.

Again, $$\frac{d^2y}{dx^2}=\frac{N}{\sqrt{2\pi}\sigma^3}\,e^{-x^2/2\sigma^2}\left(-1+\frac{x^2}{\sigma^2}\right),$$

which vanishes when $x=\pm\sigma$, and at these two points, H, H′, we

therefore have 'points of *inflexion*' where the bend of the curve changes its direction.

The axis of y about which there is symmetry evidently locates the mean error, in this case zero; in fact the mean and mode coincide, so that the mean or zero error is also the one which most frequently occurs, and any two other errors which are equal in magnitude but above and below the mean respectively occur with equal frequency: *i.e.* the frequency of positive errors is balanced by the equal frequency of negative errors on the other side of the mean, making the median error likewise zero.

Again, the area $\int_{+x_1}^{+x_2} ydx$ measures the frequency of errors lying between x_1 and x_2 above the mean; $\int_{-x}^{+x} ydx$ registers the frequency

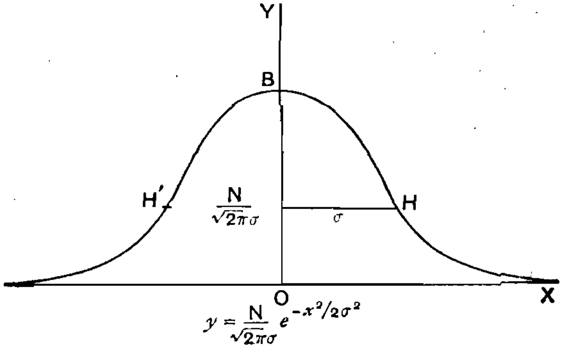


FIG. (41).

of errors between 0 and x, or of deviations up to this magnitude, on either side of the mean; and, in particular, for all errors

$$\text{the total frequency} = \int_{-\infty}^{+\infty} ydx$$

$$= \frac{N}{\sqrt{2\pi}\,.\,\sigma} \int_{-\infty}^{+\infty} e^{-x^2/2\sigma^2} dx$$

$$= \frac{N}{\sqrt{2\pi}\,.\,\sigma} (\sqrt{2\pi}\,.\,\sigma) \text{ (as on p. 206)}$$

$$= N.$$

This enables us, by means of the fundamental definition, at once to write down the probability of errors between any stated limits and explains the origin of the name, the probability curve, which

is sometimes given to the equation. Thus we have *the probability of an error between* $+x_1$ *and* $+x_2$

$$=\frac{\text{frequency of errors between the given limits}}{\text{frequency of all errors}}$$

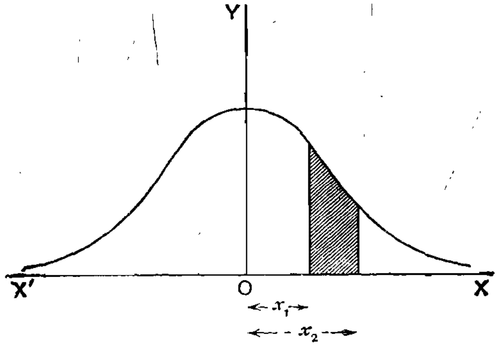
$$=\int_{x_1}^{x_2} ydx \Big/ \int_{-\infty}^{+\infty} ydx$$

$$=\frac{1}{\sqrt{2\pi}\,.\,\sigma}\int_{x_1}^{x_2} e^{-x^2/2\sigma^2}dx \quad . \quad . \quad . \quad (1)$$

Incidentally, the probability of an error between x and $(x+\delta x)$

$$=\frac{y\delta x}{N}$$

$$=\frac{\delta x}{\sqrt{2\pi}\,.\,\sigma}e^{-x^2/2\sigma^2} \quad . \quad . \quad . \quad (2)$$



$$y=\frac{N}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}$$

Fig. (42).

Geometrically, the area represented by the shaded portion of fig. (42) measures the frequency of errors between $+x_1$ and $+x_2$, while the complete area between the curve and axis X'OX measures the total frequency, so that the probability of an error between $+x_1$ and $+x_2$ is measured by the proportion which the area of the shaded portion bears to the whole area.

If in the above expression (1) we put $x/\sigma=\xi$, so that $\frac{dx}{d\xi}=\sigma$, it becomes

$$\frac{1}{\sqrt{2\pi}}\int_{\xi_1}^{\xi_2} e^{-\frac{1}{2}\xi^2}d\xi, \quad . \quad . \quad . \quad (3)$$

which is known as the *probability integral*, ξ_1 and ξ_2 being the

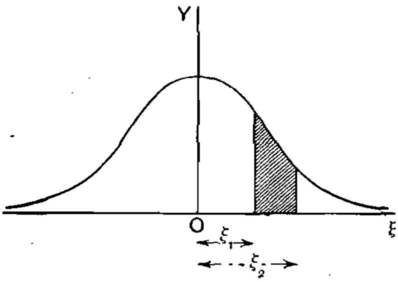
values of ξ which correspond to the values x_1 and x_2 of x. But this integral measures the area of the shaded portion of the curve

$$y=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\xi^2} \quad . \quad . \quad . \quad (4)$$

shown in fig. (43), which is really the normal curve over again, but drawn on a different scale, namely, with the ordinates reduced in the ratio N : σ and with the standard deviation σ taken as the unit of measurement for x, for $\xi=1, 2, 3$ . . . when $x=\sigma, 2\sigma, 3\sigma$, . . . This has the effect of making the total area unity and the area given by

$$\frac{1}{\sqrt{2\pi}}\int_{\xi_1}^{\xi_2} e^{-\frac{1}{2}\xi^2}d\xi \quad . \quad . \quad . \quad . \quad (3) \text{ bis}$$

now directly *measures the probability of an error between $\sigma\xi_1$ and $\sigma\xi_2$.*



$y=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\xi^2}$

FIG. (43).

Tables have been prepared (see pp. 284, 285) which enable us to write down the value of this integral for different values of ξ_1 and ξ_2 between certain limits (see Appendix, Note 10).

Let us take an example to show how the curve may be used, and we choose one leading to a binomial distribution, so giving an expression for the probability by first principles, in order to compare the two methods.

Example.—Suppose we toss simultaneously 100 coins, and suppose the chance of success, say 'heads,' is the same for each coin and equal to 1/2. In that case, according to the binomial theory,

the probability of 100 heads $=(1/2)^{100}$,

,, ,, 99 heads and 1 tail $={}^{100}C_1(1/2)^{99}(1/2)$,

,, ,, 98 heads and 2 tails $={}^{100}C_2(1/2)^{98}(1/2)^2$, and so on.

The most probable number of heads $=np=(100)(1/2)=50$. This does not mean, as explained before, that if we perform the experiment once we are sure on that one occasion to get exactly 50 heads and 50 tails, but that if we go on repeating the experiment we shall in the long run get 50 heads and 50 tails turning up more often than any other combination.

Let it be required to find the probability of getting *at least* 55

heads, that is, we want the probability of getting 55 heads or more, and this is given by

$$^{100}C_{45}(\tfrac{1}{2})^{55}(\tfrac{1}{2})^{45}+{}^{100}C_{44}(\tfrac{1}{2})^{56}(\tfrac{1}{2})^{44}+\ \ldots\ +{}^{100}C_{1}(\tfrac{1}{2})^{99}(\tfrac{1}{2})+(\tfrac{1}{2})^{100}$$

$$=\frac{1}{2^{100}}[{}^{100}C_{45}+{}^{100}C_{44}+\ \ldots\ +{}^{100}C_{1}+1],$$

a sum not very readily calculated if we have to go at it in a straightforward manner.

Now let us turn to the curve of error method. The standard deviation for the distribution is given by

$$\sigma=\sqrt{npq}=\sqrt{(100\times\tfrac{1}{2}\times\tfrac{1}{2})}=5.$$

Since the mean number of heads to be expected if the experiment is repeated a considerable number of times $=50$, we want to find the probability of an error equal to or greater than 5, *i.e.* an error lying between σ and $+\infty$, because $\sigma=5$.

But the probability of an error between $\sigma\xi_1$ and $\sigma\xi_2$

$$=\frac{1}{\sqrt{2\pi}}\int_{\xi_1}^{\xi_2}e^{-\frac{1}{2}\xi^2}d\xi,\text{ by (3) } bis.$$

Hence the required probability

$$=\frac{1}{\sqrt{2\pi}}\int_{1}^{\infty}e^{-\frac{1}{2}\xi^2}d\xi$$

$=0{\cdot}15866$, by the probability integral tables.

In other words, if we repeated the experiment 100 times, we might expect 55 or more heads about 16 times.

We can now show that *if* X_1, X_2 *are two uncorrelated variables obeying the normal law, then* $(w_1X_1+w_2X_2)$ *will obey the same law.*

Suppose x_1, x_2 are observed deviations from the mean values X_1, X_2 in one particular record, σ_1, σ_2 being the respective S.D.'s.

Let $X=w_1X_1+w_2X_2$, and let x be the deviation in X corresponding to deviations x_1, x_2 in the given variables.

Thus $$X+x=w_1(X_1+x_1)+w_2(X_2+x_2)$$
$$=(w_1X_1+w_2X_2)+(w_1x_1+w_2x_2).$$

Therefore, $$x=w_1x_1+w_2x_2.$$

But the same error x may be obtained by giving x_1, x_2 many different values provided their weighted sum is unaltered. Let us first keep x_1 constant, so that the corresponding value of x_2 required

to produce an error lying between x and $(x+\delta x)$, where δx is small, must be such that

$$x<w_1x_1+w_2x_2<x+\delta x,$$

i.e.
$$x-w_1x_1<w_2x_2<x-w_1x_1+\delta x,$$

i.e. x_2 lies between $(x-w_1x_1)/w_2$ and $(x-w_1x_1+\delta x)/w_2$, and the probability for this

$$=\frac{\delta x}{w_2}\cdot\frac{1}{\sqrt{2\pi}\cdot\sigma_2}e^{-(x-w_1x_1)^2/2\sigma^2{}_2w^2{}_2}, \text{ by (2).}$$

Now this is in a form which only involves δx, x, and x_1, and we get the total probability for an error lying between x and $(x+\delta x)$ by giving all possible values to the error x_1.

But the probability for x_1 itself to lie between x_1 and $(x_1+\delta x_1)$

$$=\frac{1}{\sqrt{2\pi}\sigma_1}\int_{x_1}^{x_1+\delta x_1}e^{-x^2/2\sigma^2{}_1}dx$$

$$=\frac{\delta x_1}{\sqrt{2\pi}\sigma_1}e^{-x^2{}_1/2\sigma^2{}_1}, \text{ by (2),}$$

and the probability for this to concur with a suitable x_2 to produce an error in the weighted sum lying between x and $(x+\delta x)$, on the assumption that X_1 and X_2 are independent, is therefore

$$=\left[\frac{\delta x_1}{\sigma_1\sqrt{2\pi}}e^{-x^2{}_1/2\sigma^2{}_1}\right]\left[\frac{\delta x}{w_2}\cdot\frac{1}{\sigma_2\sqrt{2\pi}}e^{-(x-w_1x_1)^2/2\sigma^2{}_2w^2{}_2}\right]$$

$$=\frac{\delta x}{w_2 2\pi\sigma_1\sigma_2}e^{-\frac{x^2{}_1}{2\sigma^2{}_1}-\frac{(x-w_1x_1)^2}{2\sigma^2{}_2w^2{}_2}}\delta x_1.$$

Hence the *total* probability for an error lying between x and $(x+\delta x)$ is obtained by integrating this result, that is, summing all possible probabilities, between $x_1=-\infty$ and $x_1=+\infty$. This gives

$$\frac{\delta x}{w_2\cdot 2\pi\sigma_1\sigma_2}\int_{-\infty}^{+\infty}e^{-x^2{}_1\left(\frac{1}{2\sigma^2{}_1}+\frac{w^2{}_1}{2\sigma^2{}_2w^2{}_2}\right)+2\frac{w_1x}{2\sigma^2{}_2w^2{}_2}x_1-\frac{x^2}{2\sigma^2{}_2w^2{}_2}}dx_1$$

$$=\frac{\delta x}{w_2\cdot 2\pi\sigma_1\sigma_2}\int_{-\infty}^{+\infty}e^{-x^2{}_1\frac{\sigma^2}{2\sigma^2{}_1\sigma^2{}_2w^2{}_2}+\frac{2w_1x}{2\sigma^2{}_2w^2{}_2}x_1-\frac{x^2}{2\sigma^2{}_2w^2{}_2}}dx_1,$$

where $\sigma^2=w^2{}_1\sigma^2{}_1+w^2{}_2\sigma^2{}_2$

$$=\frac{\delta x}{w_2\cdot 2\pi\sigma_1\sigma_2}\int_{-\infty}^{+\infty}e^{-\frac{1}{2}\left(\frac{\sigma x_1}{\sigma_1\sigma_2w_2}-\frac{w_1\sigma_1}{\sigma_2w_2\sigma}x\right)^2+x^2\left(\frac{\sigma^2{}_1w^2{}_1}{2\sigma^2{}_2w^2{}_2\sigma^2}-\frac{1}{2\sigma^2{}_2w^2{}_2}\right)}dx_1$$

$$=\frac{\delta x}{w_2\cdot 2\pi\sigma_1\sigma_2}e^{\frac{x^2(\sigma^2{}_1w^2{}_1-\sigma^2)}{2\sigma^2{}_2w^2{}_2\sigma^2}}\left[\int_{-\infty}^{+\infty}e^{-\xi^2}d\xi\right]\frac{\sqrt{2}\cdot\sigma_1\sigma_2w_2}{\sigma},$$

where $\xi=\frac{1}{\sqrt{2}}\left(\frac{\sigma x_1}{\sigma_1\sigma_2 w_2}-\frac{w_1\sigma_1}{\sigma_2 w_2\sigma}x\right)$

$$=\frac{\delta x}{w_2 \,.\, 2\pi\sigma_1\sigma_2}e^{-x^2 \,.\, \frac{\sigma^2_2 w^2_2}{2\sigma^2_2 w^2_2\sigma^2}}[\sqrt{\pi}]\frac{\sqrt{2}\,.\,\sigma_1\sigma_2 w_2}{\sigma}$$

$$=\frac{\delta x}{\sqrt{2\pi}\,.\,\sigma}e^{-x^2/2\sigma^2},$$

which proves that the error x obeys the normal law with

$$\text{S.D.}=\sqrt{(w^2_1\sigma^2_1+w^2_2\sigma^2_2)} \quad . \quad . \quad . \quad . \quad (5)$$

The above principle is readily extended, for if

$$X=w_1X_1+w_2X_2+ \ . \ . \ . \ +w_nX_n,$$

X_1, X_2, . . . X_n being independent variables obeying the normal law, then X also obeys the normal law and its

$$\text{S.D.}=\sqrt{(w^2_1\sigma^2_1+w^2_2\sigma^2_2+ \ . \ . \ . \ +w^2_n\sigma^2_n)} \quad . \quad . \quad (6)$$

In discussing the results of random sampling we worked upon the principle that, given a number of sample observations of any statistical constant, a mean or a percentage or a coefficient of regression or anything else, an error or deviation as large as σ, the standard deviation, from the true value for the whole population might quite likely occur, but that an error exceeding 3σ would be unlikely, and we explained that, as a result of convention, the probable error, equal to $\frac{2}{3}\sigma$ roughly, was largely used in place of σ by many writers. We have now to examine the basis of this principle, and the first point to notice is that it only strictly applies to a normal distribution.

To find the probability of an error lying between $-\sigma$ and $+\sigma$ in a normal distribution.

The required probability $=\frac{1}{\sqrt{2\pi}\,.\,\sigma}\int_{-\sigma}^{+\sigma}e^{-x^2/2\sigma^2}dx$

$$=\frac{1}{\sqrt{2\pi}}\int_{-1}^{+1}e^{-\xi^2/2}d\xi \text{ (where } x=\sigma\xi)$$

$$=\frac{2}{\sqrt{2\pi}}\int_{0}^{1}e^{-\xi^2/2}d\xi$$

$=0{\cdot}6827$, by means of the tables.

This then is the probability that the error in a given sample shall not exceed the S.D., σ. The probability that the error shall exceed

σ is accordingly $(1-0{\cdot}68)=0{\cdot}32$. It therefore appears that the odds against an error exceeding this amount are 68 to 32, or about 2 to 1.

The probability of an error between -2σ *and* $+2\sigma$

$$=\frac{1}{\sqrt{2\pi}}\int_{-2}^{+2}e^{-\xi^2/2}d\xi$$

$$=0{\cdot}9545,$$

and the probability of an error outside these limits $=0{\cdot}0455$.

Hence the odds against an error exceeding 2σ are about 21 to 1.

The probability of an error between -3σ *and* $+3\sigma$

$$=\frac{1}{\sqrt{2\pi}}\int_{-3}^{+3}e^{-\xi^2/2}d\xi$$

$$=0{\cdot}9973.$$

Hence the odds against an error exceeding 3σ are about 370 to 1.

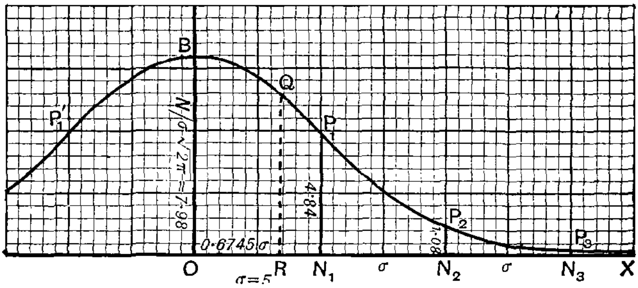


Fig. (44).

That these results are reasonable can be seen by an examination of the curve of error

$$y=\frac{N}{\sqrt{2\pi}\,.\,\sigma}e^{-x^2/2\sigma^2},$$

the graph of which is drawn, fig. (44), in the particular case when $\sigma=5$, $N=100$. The maximum ordinate is thus $=20/\sqrt{2\pi}=7{\cdot}98$, and the curve becomes

$$y=7{\cdot}98e^{-x^2/50}.$$

When $x=\sigma=5$, $y=(7{\cdot}98)(0{\cdot}606)=4{\cdot}84$, P_1N_1 in the figure.

,, $x=2\sigma=10$, $y=(7{\cdot}98)(0{\cdot}135)=1{\cdot}08$, P_2N_2 ,, ,,

,, $x=3\sigma=15$, $y=(7{\cdot}98)(0{\cdot}011)=0{\cdot}09$, P_3N_3 ,, ,,

There is a point of inflexion where the curve changes its direction at P_1, also at the companion point P'_1 on the other side of OB.

The areas ON_1P_1B, ON_2P_2B, ON_3P_3B, P_3N_3X represent respectively the frequencies of errors 0 to σ, 0 to 2σ, 0 to 3σ, 3σ and over (considering only errors on the positive side, that is, deviations above the mean), and the figure shows how very improbable is a deviation from the mean exceeding 3σ, for the area between the curve and axis beyond this limit is negligible. Put in another way, a range of 6σ should include practically all the observations in the sample.

The *probable error* has in the past received various names, such as mean error, median error, quartile deviation, and although some of these may seem more applicable and less confusing than the name to which it has settled down, there is perhaps not sufficient excuse for unsettling it again, even had we the power to do so, by attempting a return to one of these old names.

If its magnitude be r it is defined to be such that the chance of an error falling within the limits $-r$ and $+r$ is exactly equal to the chance of an error falling outside these limits, in fact it is an even chance whether a particular error falls within these limits or not.

Since area measures frequency it follows that the ordinates drawn through the probable errors divide both halves of the normal curve (above and below the mean) into two equal parts; the one above the mean, QR, is shown in fig. (44), and consequently the area OBQR = the area QRX, in that figure. These ordinates therefore coincide with the quartiles, and the probable error is precisely the same measure as the quartile deviation.

The magnitude of the error is readily calculated from the probability integral table, for, by definition, we have

$$\tfrac{1}{2}=\frac{1}{\sqrt{2\pi}\,.\,\sigma}\int_{-r}^{+r} e^{-x^2/2\sigma^2}dx$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-r/\sigma}^{+r/\sigma} e^{-\xi^2/2}d\xi \text{ (where } x=\sigma\xi).$$

Hence
$$\tfrac{1}{4}=\frac{1}{\sqrt{2\pi}}\int_{0}^{r/\sigma} e^{-\xi^2/2}d\xi,$$

and the probability integral table at once gives

$$r=0{\cdot}6745\sigma=\text{approximately } \tfrac{2}{3}\sigma.$$

Thus we have the frequently quoted rule that the

quartile deviation $=\frac{2}{3}$**(standard deviation),** . . (7)

or **probable error** $=$ **0·6745 (S.D.)**

The probability of an error lying between $-3r$ and $+3r$

$$=\frac{1}{\sqrt{2\pi}\,.\,\sigma}\int_{-3r}^{+3r}e^{-x^2/2\sigma^2}dx$$

$$=\frac{2}{\sqrt{2\pi}}\int_0^{3(0\cdot6745)}e^{-\xi^2/2}d\xi \text{ (where } x=\sigma\xi \text{ as before)}$$

$$=0\cdot9570.$$

Thus the odds against a deviation exceeding three times the probable error occurring in a single trial are about 22 to 1, or much the same as the odds against a deviation exceeding twice the S.D.

There remains one other standard of measurement in connection with errors which is at least deserving of mention, namely, what we have previously called the *mean deviation*, which may be denoted by η. It is simply the mean of all errors without regard to sign; thus, since $y\delta x$ measures the frequency of an error lying between x and $(x+\delta x)$

$$\eta=2\int_0^\infty xydx\Big/2\int_0^\infty ydx$$

$$=\int_0^\infty xe^{-x^2/2\sigma^2}dx\Big/\int_0^\infty e^{-x^2/2\sigma^2}dx$$

$$=\sigma\int_0^\infty \xi e^{-\xi^2/2}d\xi\Big/\int_0^\infty e^{-\xi^2/2}d\xi \text{ (where } x=\sigma\xi)$$

$$=\sqrt{2}\sigma\int_0^\infty te^{-t^2}dt\Big/\int_0^\infty e^{-t^2}dt \text{ (where } \xi^2=2t^2)$$

$$=\sqrt{2}\sigma\left[-\frac{e^{-t^2}}{2}\right]_0^\infty\Big/\frac{\sqrt{\pi}}{2}$$

$$=\sigma\sqrt{2/\pi}$$

$$=0\cdot7979\sigma,$$

hence the rough rule that the

mean deviation $=\frac{4}{5}$**(standard deviation)** . . . (8)

It must be borne in mind that all the above rules relating to errors—using the term as synonymous with the deviations of single or sample observations from the mean of a considerable number of the same character—strictly apply, as we said before, to the normal

curve of error and are only approximately true for other distributions, the approximation being the closer the nearer they approach to the normal form and the larger the number of observations involved. They have been tested in some cases in earlier chapters (see, for example, Chapter VII.), and the results obtained, even with very skew distributions of comparatively small numbers of observations, are at all events close enough to suggest the utility of the rules in more favourable cases.

The effect of variability on errors. The probability of an error lying between 0 and t

$$=\frac{1}{\sqrt{2\pi}\,.\,\sigma}\int_0^t e^{-x^2/2\sigma^2}dx.$$

Put $x=x'/m$, and this becomes

$$\frac{1}{\sqrt{2\pi}\,.\,\sigma}\int_0^{mt} e^{-x'^2/2\sigma^2m^2}\frac{dx'}{m}$$

$$=\frac{1}{\sqrt{2\pi}\,.\,(m\sigma)}\int_0^{(mt)} e^{-x^2/2(m\sigma)^2}dx.$$

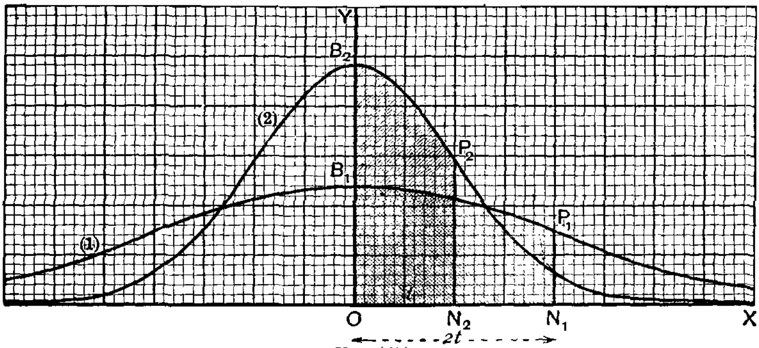


FIG. (45).

Thus, if the variability be increased m-fold the range of error (of equal probability) is increased m-fold, so that if we have two sets of N observations, with the variability of one set double that of the other, the range of error also in the one set is double that which is equally likely to occur in the other. This is brought out fairly clearly in fig. (45), which is the result of plotting the curve

$$y=\frac{N}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}$$

in the two cases. The variability σ of curve (1) is double that of curve (2); if then we measure along OX in the figure

$$ON_1 = 2ON_2 = 2t,$$

the area $B_1ON_1P_1$ will be equal to the area $B_2ON_2P_2$, showing that the probability for an error between 0 and $2t$ in the one case is equal to the probability for an error between 0 and t in the other case.

[James Bernoulli (1654-1705), the eldest of three remarkable brothers, showed how the binomial theorem could be used to estimate the probability that the ratio of the number of successes to the number of failures under defined conditions should lie between set limits, where *success* means that a certain event happens and *failure* means that it fails to happen.

It was Gauss who first actually published a proof (1809) of the equation of the normal curve, although Laplace had suggested as early as 1783 the utility of a probability integral table, $\int e^{-t^2}dt$. Gauss's proof depended upon certain axioms which cannot be established and are not necessarily true, one of which was that 'errors above and below the mean are equally probable.' Laplace and Poisson improved upon Gauss and succeeded without assuming this axiom, but with the aid of theorems due to Euler and Stirling, in developing the continuous probability integral from the discontinuous binomial series.

Further extensions of the normal curve applicable to skew distributions have been worked out by other writers, such as Galton and McAlister, Fechner, Lipps, Werner, Charlier, Kapteyn, and finally by Edgeworth, who has contributed materially to the development of the idea of 'the Law of Great Numbers.' Karl Pearson approaching the subject of skew variation from the same point but by an original route, has discovered a complete system of curves suitable for fitting almost all kinds of distributions in homogeneous material, especially such as are met with in the biological world.

(See Todhunter, *History of Probability*.
Edgeworth, *Law of Error* in the *Encyclopaedia Britannica* (10th edition).
Pearson, *Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson: A Rejoinder; Biometrika*, vol. iv., pp. 169-212).]

CHAPTER XIX

FREQUENCY SURFACE FOR TWO CORRELATED VARIABLES

It may serve at this stage to widen the outlook upon the subject of correlation for those who are able to follow it up on mathematical lines if we briefly consider the algebraical expression for the combined distribution of two variables.

Let the variables be X_1, X_2. They may be absolutely independent or they may be related in some way, but in either case we shall assume it possible to set up a one-to-one correspondence between them: thus, X_1 might represent the marriage rate and X_2 the index number for wholesale prices, and we might always pair together the X_1 and the X_2 which refer to the same year, as in the correlation example in a previous chapter; moreover this pairing might still be effected even if there were really no other connection at all between X_1 and X_2.

If then x_1, x_2 typify the deviations of X_1, X_2 from their respective means (the means in the above case being derived by averaging the figures for a number of years), it is possible to write down an expression of the form

$$y = F(x_1, x_2)$$

for determining the probability of deviations between x_1 and $(x_1 + \delta x_1)$, x_2 and $(x_2 + \delta x_2)$, occurring simultaneously (in the same year, in the above case); or, to put the same thing in another way, $y \delta x_1 \delta x_2$ would represent the proportional frequency with which such deviations might be expected to occur together in a large number of observations.

The frequency curve $y = f(x)$, where $y \delta x$ denotes the frequency with which a variable with deviation lying between x and $(x + \delta x)$ from its mean value is observed in a given distribution, was represented by plotting corresponding pairs of values of x and y as points in a plane. In the expression $y = F(x_1, x_2)$, however, we have three variables to consider, x_1 and x_2, and y which measures the frequency of the simultaneous appearance of x_1 and x_2. Such a trio may geometrically be represented by a point P (x_1, x_2, y) in

space of three dimensions, for (x_1, x_2) can first be located as a point in a fixed plane and a height y may then be measured above this plane as in fig. (46). Clearly as x_1 and x_2 vary, y also varies, and consequently the point P moves about in space, but it moves always in obedience to the relation

$$y=\mathrm{F}(x_1, x_2).$$

This relation is called the equation of the surface along which P travels, showing that it holds good for the co-ordinates (x_1, x_2, y) of any position which the point can take up on that surface. It is convenient, however, to use the notation

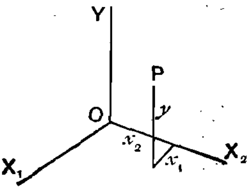


Fig. (46).

$$z=\mathrm{F}(x, y)$$

in preference to $y=\mathrm{F}(x_1, x_2)$ for the 'frequency surface,' because OX, OY are nearly always taken to represent the axes of reference in space of two dimensions (*i.e.* in a plane), and by a natural extension OX, OY, OZ are taken to represent the axes of reference in space of three dimensions, fig. (47).

We proceed to discuss the frequency surface for two variables, and we shall start with the comparatively simple case when the variables are completely independent.

Frequency surface showing distribution of two completely independent variables each subject to the normal law.

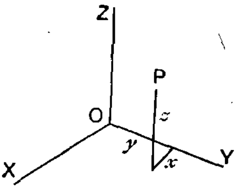


Fig. (47).

Let X, Y be the variables, and let x, y denote deviations from their means $\bar{\mathrm{X}}$, $\bar{\mathrm{Y}}$, the point $(\bar{\mathrm{X}}, \bar{\mathrm{Y}})$ being taken as origin of co-ordinates and the usual notation being adopted.

Thus the probability of a deviation between x and $(x+\delta x)$ occurring

$$=\frac{\delta x}{\sqrt{2\pi}\,.\,\sigma_x}e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}},$$

and the probability of a deviation between y and $(y+\delta y)$ occurring

$$=\frac{\delta y}{\sqrt{2\pi}\,.\,\sigma_y}e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}}.$$

Therefore the probability of such deviations occurring together since the variables are supposed completely independent

$$=\left(\frac{\delta x}{\sqrt{2\pi}\,.\,\sigma_x}e^{-x^2/2\sigma_x^2}\right)\left(\frac{\delta y}{\sqrt{2\pi}\,.\,\sigma_y}e^{-y^2/2\sigma_y^2}\right)$$

$$=\frac{\delta x\delta y}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}.$$

Hence the frequency with which such pairs of deviations are observed together if n be the total number of observations

$$=n/2\pi\sigma_x\sigma_y \cdot e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}\delta x\delta y.$$

Denoting this by $z\delta x\delta y$, we get for the required frequency surface,

$$z=n/2\pi\sigma_x\sigma_y \cdot e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} \quad . \quad . \quad . \quad (1)$$

If we give y some particular value, y_1, we find from the above equation that the law of frequency for the corresponding x is

$$z=\frac{n}{2\pi\sigma_x\sigma_y} \cdot e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y_1^2}{\sigma_y^2}\right)}$$

$$=\left[\frac{n}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\frac{y_1^2}{\sigma_y^2}}\right]e^{-x^2/2\sigma_x^2}$$

$$=\frac{n_1}{\sqrt{2\pi}\cdot\sigma_x}e^{-x^2/2\sigma_x^2},$$

where n_1 has been written in place of

$$\left(\frac{n}{\sqrt{2\pi}\cdot\sigma_y}e^{-y_1^2/2\sigma_y^2}\right).$$

But this is evidently a normal curve in the plane X_1OZ_1, having the same mean, $\bar{X}$, and the same S.D., σ_x, whatever be the value of y_1.

Hence all arrays of X are similar, having the same mean and the same standard deviation, and this, by symmetry, also applies to all arrays of y.

Now put z equal to some constant, k, in equation (1), so that

$$k=\frac{n}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)};$$

$$\therefore \frac{2\pi\sigma_x\sigma_y}{n}(k)=e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}.$$

Since the left-hand side of this equation is constant for different values of (x, y), it follows that the right-hand side is also constant, and hence

$$\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}=c, \quad . \quad . \quad . \quad (2)$$

where c is a constant.

We conclude that the values of x and y which can occur together with a given frequency, k, are such that the point (x, y) always lies

somewhere on the ellipse (2) in the plane $z=k$, fig. (48); *e.g.* values in the neighbourhood of x_1 and y_1 occur with the same frequency as values in the neighbourhood of x_2 and 0, because in the figure the points (x_1, y_1, k) and $(x_2, 0, k)$ both lie on the ellipse defined by

$$z=k, \quad \frac{x^2}{\sigma_x^{\,2}}+\frac{y^2}{\sigma_y^{\,2}}=c.$$

The different ellipses which can be obtained by varying the frequency, and consequently varying c, are clearly concentric, similar, and similarly situated if they are orthogonally projected on to the plane $z=0$, for the effect of such projection is that any

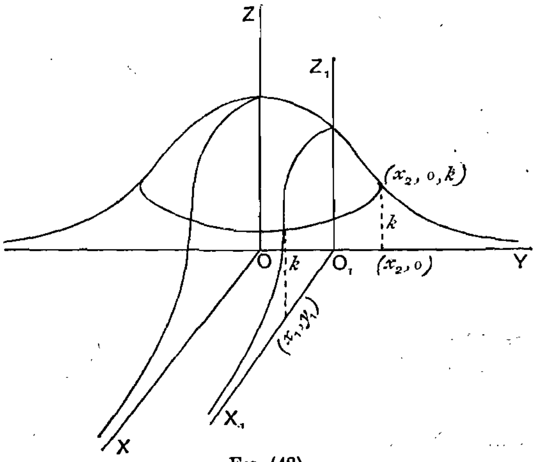


FIG. (48).

point (x, y, z) drops down on to the point $(x, y, 0)$ which stands immediately below it in the plane XOY.

The general shape of the surface can be gathered from fig. (48) where the ellipse in the plane $z=k$, and the normal curves in the planes $x=0$, $y=0$, and $y=y_1$ have been drawn.

It will also be noted that if the scales of x and y are altered by writing $\dfrac{x}{\sigma_x}=x'$ and $\dfrac{y}{\sigma_y}=y'$, so that unit change in each may be the same, the ellipse (2) becomes a circle

$$x'^2+y'^2=c.$$

This change of scales is equivalent geometrically to projecting orthogonally the ellipse into a circle; of course the planes of projection are not the same as in the previous orthogonal projection mentioned.

Frequency surface for two correlated variables. Let the variables be X and Y, and let us work as before with their deviations x and y, which is equivalent to taking the mean point $(\bar{X}, \bar{Y})$ of all the observations as origin.

Now the line of regression giving the best y, or the y of greatest frequency, corresponding to any x is

$$y = r\frac{\sigma_y}{\sigma_x}x,$$

with the usual notation, r being the coefficient of correlation between X and Y.

Hence the error made in estimating any y from this equation instead of taking the y given by observation is

$$\eta = y \text{ (observed) } - y \text{ (estimated)}$$

$$= y - r\frac{\sigma_y}{\sigma_x}x. \quad \text{[See fig. (49).]}$$

Thus, corresponding to every pair of observations (x, y) there is an η, and the same η will be repeated just as often as the same pair of observations (x, y) is repeated.

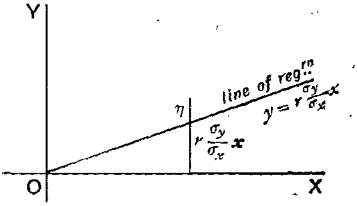


FIG. (49)

Therefore the frequency distribution of (x, η) must exactly correspond to that of (x, y).

Further, the correlation of the variables x and η is zero, for positive and negative errors η are equally likely to occur for different values of x; in fact, this coefficient of correlation is $\Sigma(x\eta)/n\sigma_x\sigma_\eta$, and

$$\Sigma(x\eta) = \Sigma\left[x\left(y - r\frac{\sigma_y}{\sigma_x}x\right)\right]$$

$$= \Sigma(xy) - r \cdot \frac{\sigma_y}{\sigma_x} \cdot \Sigma(x^2)$$

$$= np - \frac{p}{\sigma_x\sigma_y} \cdot \frac{\sigma_y}{\sigma_x} \cdot n\sigma_x^2$$

$$= np - np$$

$$= 0.$$

Assuming then that the variables x and η are quite independent, the probability of them occurring together is readily written down, for it is simply the product of their separate probabilities.

But the probability of a deviation between x and $(x+\delta x)$ occurring, if we consider this variable alone, is

$$\frac{\delta x}{\sqrt{2\pi}\sigma_x}e^{-\frac{x^2}{2\sigma_x^2}},$$

and the probability of a deviation between η and $(\eta+\delta\eta)$ occurring, if we consider this variable alone, is

$$\frac{\delta\eta}{\sqrt{2\pi}\sigma_\eta}e^{-\frac{\eta^2}{2\sigma_\eta^2}}.$$

Hence the probability of a combined occurrence of such deviations

$$=\left(\frac{\delta x}{\sqrt{2\pi}\sigma_x}e^{-\frac{x^2}{2\sigma_x^2}}\right)\left(\frac{\delta\eta}{\sqrt{2\pi}\sigma_\eta}e^{-\frac{\eta^2}{2\sigma_\eta^2}}\right)$$

$$=\frac{\delta x\delta\eta}{2\pi\sigma_x\sigma_\eta}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{\eta^2}{\sigma_\eta^2}\right)}$$

$$=\frac{\delta x\delta\eta}{2\pi\sigma_x\sigma_\eta}e^{-\frac{1}{2}\left\{\frac{x^2}{\sigma_x^2}+\frac{\left(y-r\frac{\sigma_y}{\sigma_x}x\right)^2}{\sigma_\eta^2}\right\}}$$

$$=\frac{\delta x\delta\eta}{2\pi\sigma_x\sigma_\eta}e^{-\frac{1}{2}\left\{\frac{y^2}{\sigma_\eta^2}-2xy\,.\,\frac{r\sigma_y}{\sigma_x\sigma_\eta^2}+x^2\left(\frac{1}{\sigma_x^2}+\frac{r^2\sigma_y^2}{\sigma_x^2\sigma_\eta^2}\right)\right\}}.$$

But $$n\sigma_\eta^2=\Sigma\left(y-r\frac{\sigma_y}{\sigma_x}x\right)^2$$

$$=\Sigma(y^2)-2r\,.\,\frac{\sigma_y}{\sigma_x}\,.\,\Sigma(xy)+r^2\frac{\sigma_y^2}{\sigma_x^2}\Sigma(x^2)$$

$$=n\sigma_y^2-2r\,.\,\frac{\sigma_y}{\sigma_x}\,.\,n\sigma_x\sigma_y r+r^2\frac{\sigma_y^2}{\sigma_x^2}\,.\,n\sigma_x^2$$

$$=n\sigma_y^2(1-r^2).$$

Similarly, $$n\sigma_\xi^2=n\sigma_x^2(1-r^2),$$

where ξ is the error made in estimating x from $x=r\frac{\sigma_x}{\sigma_y}y$;

$$\therefore\ \frac{\sigma_\xi^2}{\sigma_x^2}=\frac{\sigma_\eta^2}{\sigma_y^2}=(1-r^2).$$

Thus $$\frac{\sigma_y}{\sigma_x\sigma_\eta^2}=\frac{1}{\sigma_x\sigma_\eta}\,.\,\frac{\sigma_y}{\sigma_\eta}=\frac{1}{\sigma_x\sigma_\eta}\,.\,\frac{\sigma_x}{\sigma_\xi}=\frac{1}{\sigma_\xi\sigma_\eta},$$

and $$\left(\frac{1}{\sigma_x^2}+\frac{r^2\sigma_y^2}{\sigma_x^2\sigma_\eta^2}\right)=\frac{1}{\sigma_x^2}\left(1+r^2\cdot\frac{\sigma_y^2}{\sigma_\eta^2}\right)$$
$$=\frac{1}{\sigma_x^2}\left(1+r^2\cdot\frac{1}{1-r^2}\right)$$
$$=\frac{1}{\sigma_x^2(1-r^2)}$$
$$=\frac{1}{\sigma_\xi^2}.$$

Hence the probability of the combined occurrence of deviations x to $(x+\delta x)$, η to $(\eta+\delta\eta)$

$$=\frac{\delta x\cdot\delta\eta}{2\pi\sigma_x\cdot\sigma_y\sqrt{1-r^2}}e^{-\frac{1}{2}\left\{\frac{y^2}{\sigma_\eta^2}-2xy\cdot r\cdot\frac{1}{\sigma_\xi\sigma_\eta}+x^2\cdot\frac{1}{\sigma_\xi^2}\right\}};$$

thus, if we denote by $z\delta x\delta y$ the frequency of the combined occurrence of deviations x to $(x+\delta x)$, y to $(y+\delta y)$, when n is the total number of observations, we have *

$$z=\frac{n}{2\pi\sqrt{1-r^2}\cdot\sigma_x\sigma_y}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}\right)\frac{1}{1-r^2}}.$$

When the variables X and Y are completely independent, so that r is zero, this reduces, as it should, to our previous result

$$z=\frac{n}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}.$$

In the surface $$z=\mu\cdot e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}\right)\frac{1}{1-r^2}}\quad . \quad . \quad . \quad (3)$$

where $\mu=\frac{n}{2\pi\sqrt{1-r^2}\cdot\sigma_x\sigma_y}$, if we give y some particular value y_1, we find that the law of frequency for the corresponding x is

$$z=\mu\cdot e^{-\frac{1}{2(1-r^2)}\left(\frac{y_1^2}{\sigma_y^2}+\frac{x^2}{\sigma_x^2}-2r\frac{xy_1}{\sigma_x\sigma_y}\right)}$$
$$=\mu\cdot e^{-\frac{1}{2(1-r^2)}\left\{\frac{y_1^2}{\sigma_y^2}(1-r^2)+\left(\frac{x}{\sigma_x}-r\frac{y_1}{\sigma_y}\right)^2\right\}}$$
$$=\mu\cdot e^{-\frac{y_1^2}{2\sigma_y^2}}e^{-\frac{1}{2(1-r^2)}\left(\frac{x}{\sigma_x}-r\frac{y_1}{\sigma_y}\right)^2}\quad . \quad . \quad . \quad (4)$$

[* For an outline of Karl Pearson's method of reaching the Law of Frequency for two correlated variables, and certain deductions from it, see Appendix, Note 11.]

But just as $y=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma_x^2}}$

represents exactly the same normal curve as

$$y=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}},$$

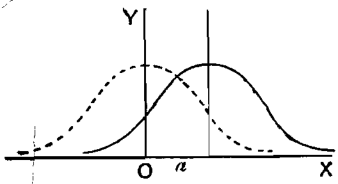


FIG. (50).

shifted through a distance a along the axis of x, fig. (50), so we conclude that the curve (4) in x and z, in the plane $y=y_1$, is exactly the same as the normal curve

$$z=\frac{\mu}{e^{y^2_1/2\sigma_y^2}}e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2(1-r^2)}},$$

shifted through a distance $ry_1\frac{\sigma_x}{\sigma_y}$ along an axis parallel to OX. In fact (4) represents a normal distribution for x, the mean, corresponding to greatest frequency when $z=\frac{\mu}{e^{y^2_1/2\sigma_y^2}}$, being determined by the intersection with the surface (3) of the planes

$$y=y_1,\ \frac{x}{\sigma_x}=r\frac{y}{\sigma_y},$$

and the standard deviation being $\sigma_x\sqrt{1-r^2}$, which we note is independent of y_1, fig. (51). To put the same thing in another way, the array of x's corresponding to a particular value y_1 of y have a mean deviating from $\bar{X}$ by $r\frac{\sigma_x}{\sigma_y}\,.\,y_1$, and a standard deviation $\sigma_x\sqrt{1-r^2}$.

In particular, when $y=0$, $z=\mu e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2(1-r^2)}}$, a normal distribution for x, the mean, corresponding to greatest frequency with $z=\mu$, being determined by the intersection with the surface (3) of the planes $y=0$, $\frac{x}{\sigma_x}=r\frac{y}{\sigma_y}$, and the standard deviation being $\sigma_x\sqrt{1-r^2}$ as before.

Similarly, when $x=x_1$, we get as in (4) a normal distribution for y,

$$z=\mu e^{-\frac{x^2_1}{2\sigma_x^2}}e^{-\frac{1}{2(1-r^2)}\left(\frac{y}{\sigma_y}-r\frac{x_1}{\sigma_x}\right)^2},$$

the mean, corresponding to greatest frequency when $z=\frac{\mu}{e^{x^2_1/2\sigma_x^2}}$, being determined by the intersection with the surface (3) of the planes

$$x=x_1,\ \frac{y}{\sigma_y}=r\frac{x}{\sigma_x},$$

and the standard deviation being $\sigma_y\sqrt{1-r^2}$, which is independent of x_1. In other words, the array of y's corresponding to a particular value x_1 of x have a mean deviating from $\bar{\bar{Y}}$ by $r\frac{\sigma_y}{\sigma_x}x_1$, and a standard deviation $\sigma_y\sqrt{1-r^2}$.

In particular, when $x=0$, $z=\mu e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2(1-r^2)}}$, a normal distribution for y, the mean, corresponding to greatest frequency with $z=\mu$, being determined by the intersection with the surface (3) of the planes $x=0$, $\frac{y}{\sigma_y}=r\frac{x}{\sigma_x}$, and the standard deviation being $\sigma_y\sqrt{1-r^2}$.

By putting z=some constant, k, and arguing just as we did in the

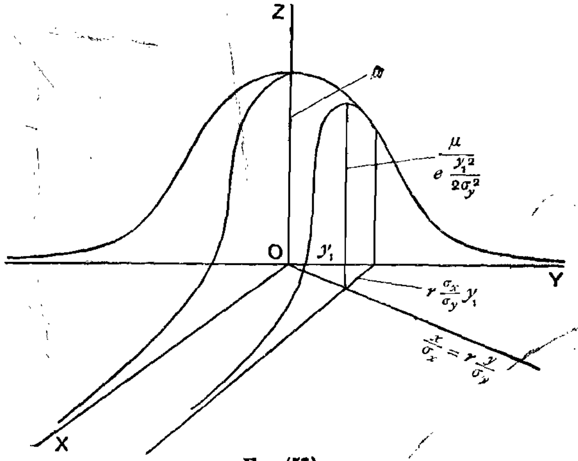


FIG. (51).

case of two independent variables, we find that all values of x and y which occur together with the same frequency define points (x, y) which lie on the ellipse

$$z=k, \quad \frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}=c.$$

The different ellipses which can be obtained by varying the frequency, and consequently varying c, are concentric, similar, and similarly situated, if they are orthogonally projected on to the plane $z=0$. The planes giving the means of the x's, or the most frequent x's, corresponding to particular values of y, and the means of the y's, or the most frequent y's, corresponding to particular values of x, meet $z=0$ in the lines of regression

$$\frac{x}{\sigma_x}=r\frac{y}{\sigma_y}, \quad \frac{y}{\sigma_y}=r\frac{x}{\sigma_x}.$$

If we alter the scales of x and y by writing $\frac{x}{\sigma_x}=x'$ and $\frac{y}{\sigma_y}=y'$ so that unit change in each shall be of the same magnitude, the frequency surface takes the form

$$z=\mu e^{-\frac{1}{2(1-r^2)}(x'^2+y'^2-2rx'y')}.$$

When $y'=0$, $z=\mu e^{-\frac{1}{2(1-r^2)}x'^2}$, a normal distribution, the mean being on the plane $x'=ry'$, and the standard deviation being $\sqrt{1-r^2}$. Similarly for $x'=0$. When $y'=y'_1$, $z=\mu e^{-\frac{1}{2}y'^2_1} e^{-\frac{1}{2(1-r^2)}(x'-ry'_1)^2}$, a normal distribution, the mean being on the plane $x'=ry'$, and the standard deviation being $\sqrt{1-r^2}$ as before. Similarly for $x'=x'_1$.

Again the ellipse which is the locus of the points $(x'y')$ obtained by putting $z=$constant, k, corresponding to variables which occur with the same frequency, is (in the plane $z=k$) now

$$x'^2+y'^2-2rx'y'=c,$$

and, projecting on to the plane $z=0$, the lines of regression are

$$x'=ry',\ y'=rx'.$$

These lines are the intersections with $z=0$ of the planes containing the means of the x''s, or the most frequent x''s, corresponding to particular y''s, and *vice versa.*

Since, geometrically, the transformation $\frac{x}{\sigma_x}=x'$, $\frac{y}{\sigma_y}=y'$, is equivalent to an orthogonal projection, we may learn something about the more general ellipse by considering properties of the simpler projected curve which are not changed by projection.

Let us first, however, find the magnitude and direction of the axes of

$$x'^2+y'^2-2rx'y'=c.$$

By turning the axes through some angle θ this equation is reducible to the form

$$\frac{x''^2}{a^2}+\frac{y''^2}{b^2}=1,$$

which is the ordinary form for an ellipse when its axes lie along the axes of co-ordinates. But the equation in x', y' is clearly symmetrical about the lines $y'=x'$ and $y'=-x'$, because y' and x' or y' and $-x'$ can be interchanged without the equation being affected. Hence these lines must give the directions of the major and minor axes.

To turn the axes of co-ordinates through an angle of 45°, fig. (52), we must write

$$x' = x'' \cos 45° - y'' \sin 45° = \frac{x'' - y''}{\sqrt{2}}$$

$$y' = x'' \sin 45° + y'' \cos 45° = \frac{x'' + y''}{\sqrt{2}}.$$

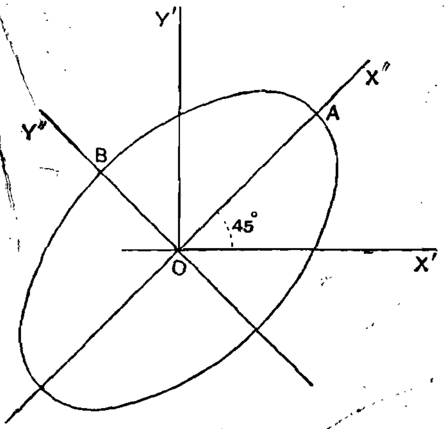


Fig. (52).

The equation of the ellipse thus becomes

$$\frac{(x''-y'')^2}{2} + \frac{(x''+y'')^2}{2} - 2r\frac{(x''-y'')(x''+y'')}{\sqrt{2}\sqrt{2}} = c,$$

i.e. $$x''^2 + y''^2 - r(x''^2 - y''^2) = c,$$

i.e. $$x''^2(1-r) + y''^2(1+r) = c,$$

i.e. $$\frac{x''^2}{\frac{c}{1-r}} + \frac{y''^2}{\frac{c}{1+r}} = 1.$$

Hence the semi-major axis is $a = \sqrt{\frac{c}{1-r}}$, and the semi-minor axis is $b = \sqrt{\frac{c}{1+r}}$. We note that as r increases from 0 to 1, a increases from $\sqrt{c}$ to ∞, while b decreases from $\sqrt{c}$ to $\sqrt{\frac{c}{2}}$. Also, as r decreases from 0 to -1, a decreases from $\sqrt{c}$ to $\sqrt{\frac{c}{2}}$, while b increases from $\sqrt{c}$ to ∞.

The ellipses, $x'^2+y'^2-2rx'y'=c$, corresponding to different values of r all pass through the points of intersection of

$$x'^2+y'^2=c \text{ and } x'y'=0.$$

But $x'^2+y'^2=c$ is what the equation of the ellipse becomes when r, the coefficient of correlation, vanishes. The connection between these curves is shown in fig. (53), which represents their projection on to the plane $z=0$. A positive correlation between x and y might be expected to increase the y corresponding to a particular positive x, if the frequency be fixed beforehand, and that is the effect which the figure also would suggest.

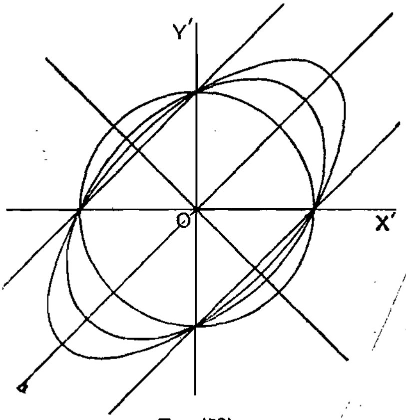


Fig. (53).

Now, in $$x'^2+y'^2-2rx'y'=c,$$
the lines of regression are

$$y'=rx',\ y'=\frac{1}{r}x',$$

and the axes of the ellipse are

$$y'=x',\ y'=-x'.$$

Hence the lines of regression are equally inclined to the axes of the ellipse as well as to the axes of co-ordinates, fig. (54).

Further, the pair of lines

$$y'=x',\ y'=-x'$$

form a harmonic pencil with the pair

$$x'=0,\ y'=0,$$

and also with the pair

$$y'=rx',\ y'=\frac{1}{r}x'.$$

This is obvious from fig. (54).

Now project back to the ellipse

$$\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}=\text{constant}.$$

The algebraical transformation for this is merely

$$x'=\frac{x}{\sigma_x},\quad y'=\frac{y}{\sigma_y}.$$

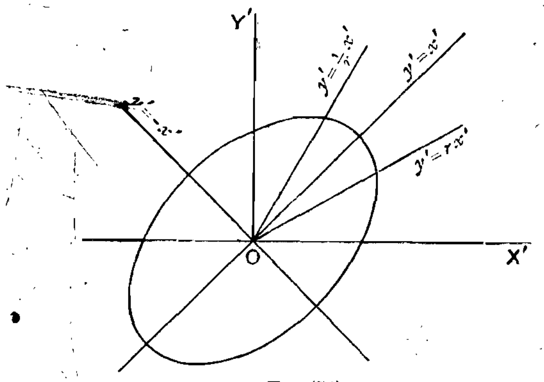


FIG. (54).

Since the harmonic property is unaltered by projection we then have the pair of lines

$$\frac{y}{\sigma_y}=\frac{x}{\sigma_x},\quad \frac{y}{\sigma_y}=-\frac{x}{\sigma_x}$$

harmonic with the pair

$$x=0,\ y=0,$$

and also with the pair

$$\frac{y}{\sigma_y}=r\frac{x}{\sigma_x},\quad \frac{y}{\sigma_y}=\frac{1}{r}\cdot\frac{x}{\sigma_x}.$$

Hence the two lines of regression corresponding to maximum correlation ($r=+1$ and $r=-1$) are harmonic with

(1) the axes of co-ordinates;

(2) the lines of regression for any r.

Again it may be easily seen that the lines

$$y'=rx' \text{ and } x'=0$$

are conjugate diameters of the ellipse

$$x'^2+y'^2-2rx'y'=c, \qquad \text{(5)}$$

for they may be written as one equation thus:

$$rx'^2-x'y'=0,$$

and this represents a pair of lines harmonic with the (imaginary) asymptotes of (5), namely, with

$$x'^2+y'^2-2rx'y'=0.$$

[The criterion for $\quad ax^2+2hxy+by^2=0$

to be harmonic with $\quad a'x^2+2h'xy+b'y^2=0$

is $\quad ab'+ba'=2hh'$.]

But it is a well-known property of conics that any pair of lines harmonic with the asymptotes are conjugate diameters of the conic.

Similarly it may be shown that the lines

$$y'=\frac{1}{r}x' \text{ and } y'=0$$

are conjugate diameters of the ellipse (5).

But, on projection, the conjugate property also is unaltered.

Hence the lines $\quad \dfrac{y}{\sigma_y}=r\dfrac{x}{\sigma_x},\ x=0,$

and the lines $\quad \dfrac{y}{\sigma_y}=\dfrac{1}{r}\dfrac{x}{\sigma_x},\ y=0$

are conjugate pairs of diameters of the ellipse

$$\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2r\frac{xy}{\sigma_x\sigma_y}=c.$$

But for conjugate diameters the midpoints of all chords parallel to either lie on the other.

Thus we come back again by another route to the familiar line of regression theorems that, for a given r, all arrays parallel to $x=0$ have their means on $\dfrac{y}{\sigma_y}=r\dfrac{x}{\sigma_x}$, and all arrays parallel to $y=0$ have their means on $\dfrac{x}{\sigma_x}=r\dfrac{y}{\sigma_y}$.

APPENDIX

1. Compound Interest Law. If the capital increases continuously, instead of going up by jumps at the end of stated periods, the connection between the original principal S_0, the rate per cent. per annum r, and the amount S_t at the end of t years is given by

$$S_t = S_0 e^{rt/100},$$

for the rate of increase is measured by

$$\frac{dS}{dt} = \frac{rS}{100},$$

which leads at once to the above equation on integrating.

Other instances of the same law are :—

(1) *A particle moving against a resistance proportional to its velocity*,

$$v_t = v_0 e^{-ct},$$

where v_t is the velocity at time t, v_0 is the original velocity, and c is some constant.

(2) *The variation of the pressure of the atmosphere with height*,

$$p_h = p_0 e^{-ch},$$

where p_h is the pressure at height h above a surface level, p_0 is the pressure at the surface, and c is some constant.

(3) *The rate of cooling*, $\theta_t = \theta_0 e^{-ct}$,

where θ_t is the excess of temperature at time t of the hot body over that of surrounding bodies, θ_0 is the excess when the measurement begins, and c is some constant.

2. Weighted Mean. Let the observations be represented by th different values, $x_1, x_2, \ldots x_n$, of the variable concerned, and le the respective weights attached to these observations be $f_1, f_2, \ldots f$ ... average, by definition,

Now, suppose a different set of weights be chosen, namely. $f'_1, f'_2, \ldots f'_n$, giving a new average

$$=\frac{x_1f'_1+x_2f'_2+\ \ldots\ +x_nf'_n}{f'_1+f'_2+\ \ldots\ +f'_n}.$$

The difference between these two expressions

$$=\frac{x_1f_1+x_2f_2+\ \ldots}{f_1+f_2+\ \ldots}-\frac{x_1f'_1+x_2f'_2+\ \ldots}{f'_1+f'_2+\ \ldots}$$

$$=\frac{(f'_1+f'_2+\ldots)(x_1f_1+x_2f_2+\ldots)-(f_1+f_2+\ldots)(x_1f'_1+x_2f'_2+\ldots)}{(f_1+f_2+\ \ldots)(f'_1+f'_2+\ \ldots)}$$

$$=\frac{\{(f_1f'_2(x_1-x_2)-f_2f'_1(x_1-x_2)\}+\{f_1f'_3(x_1-x_3)-f_3f'_1(x_1-x_3)\}+\ \ldots}{(f_1+f_2+\ \ldots)(f'_1+f'_2+\ \ldots)}$$

$$=\frac{f'_1f'_2(x_1-x_2)\left(\frac{f_1}{f'_1}-\frac{f_2}{f'_2}\right)+f'_1f'_3(x_1-x_3)\left(\frac{f_1}{f'_1}-\frac{f_3}{f'_3}\right)+\ \ldots}{(f_1+f_2+\ \ldots)(f'_1+f'_2+\ \ldots)}$$

Hence this difference is very small and the averages are very nearly equal if the weights $f_1, f_2, f_3 \ldots$ are replaced by others $f'_1, f'_2, f'_3 \ldots$ very nearly proportional to them, so that f_1/f'_1, f_2/f'_2, $f_3/f'_3 \ldots$ are not far from equality, and this is the more pronounced if the observations $x_1, x_2, x_3 \ldots$ themselves are all of the same order of magnitude and the sums of their weights, Σf and $\Sigma f'$, are large so that the expressions of type$(x_1-x_2)/(\Sigma f)(\Sigma f')$ are small.

3. Geometric and Harmonic Means. Given n numbers

$$a, b, c \ldots$$

their geometric mean, g, is defined by the formula

$$g=\sqrt[n]{(abc\ \ldots\ )},$$

and their harmonic mean, h, is defined by

$$\frac{n}{h}=\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\ \ldots$$

We note that when $a=b=c=\ \ldots\ =k$, say,

then $g=\sqrt[n]{(kkk\ \ldots)}=\sqrt[n]{(k^n)}=k,$

and $\frac{n}{h}=\frac{1}{k}+\frac{1}{k}+\frac{1}{k}+\ \ldots\ =\frac{n}{k}$

so that $h=k.$

It is worthy of remark that if the geometric mean be adopted as average in discussing the index numbers of prices it possesses an interesting property which does not hold for any of the other means in common use.

Suppose the prices of n standard commodities at three successive dates be represented by $(a_1, b_1, c_1 \ldots)$, $(a_2, b_2, c_2 \ldots)$, $(a_3, b_3, c_3 \ldots)$. Then the index numbers of the separate commodity prices at the third date, taking the prices at the first date as standard, are

$$100\frac{a_3}{a_1},\ 100\frac{b_3}{b_1},\ 100\frac{c_3}{c_1} \ldots$$

Hence the geometric mean of these n index numbers together

$$=\sqrt[n]{\left(100\frac{a_3}{a_1}\times 100\frac{b_3}{b_1}\times 100\frac{c_3}{c_1}\times \ldots\right)}$$
$$=100\sqrt[n]{(a_3b_3c_3 \ldots)}/\sqrt[n]{(a_1b_1c_1 \ldots)}$$
$$=100g_3/g_1,$$

where g_1, g_3 denote the geometric means of the n prices at the two dates.

It follows that the ratio

$$\frac{\text{index number of prices at 3rd date with prices at 1st date as standard}}{\text{index number of prices at 2nd date with prices at 1st date as standard}}$$
$$=\frac{100g_3/g_1}{100g_2/g_1}$$
$$=g_3/g_2.$$

It is therefore quite *independent of the particular date chosen as standard.*

4. **The Mean of Combined Sets of Observations.** (1) Suppose one variable x is expressed as the sum of a number of other variables, thus

$$x=a+b+c+ \ldots,$$

and suppose that we have n different values of the variables, giving equations of the type

$$x_1=a_1+b_1+c_1+ \ldots$$
$$x_2=a_2+b_2+c_2+ \ldots$$
$$\ldots \quad \ldots \quad \ldots$$
$$x_n=a_n+b_n+c_n+ \ldots$$

Hence, by addition,

$$x_1+x_2+\ldots+x_n=(a_1+\ldots+a_n)+(b_1+\ldots+b_n)+(c_1+\ldots+c_n)+\ldots$$

so that

$$n\bar{x}=n\bar{a}+n\bar{b}+n\bar{c}+\ldots$$

$$\bar{x}=\bar{a}+\bar{b}+\bar{c}+\ldots,$$

where $\bar{x}$, $\bar{a}$, $\bar{b}$ . . . denote the means of the n values of the respective variables.

Thus *the mean of a sum equals the sum of the means*, and, if some of the positive signs in $(a+b+c+\ldots)$ are made negative, there will evidently be a corresponding change of sign in $(\bar{a}+\bar{b}+\ldots)$.

Example.—Suppose 100 family budgets are collected and the items in each are separated under five heads—rent, food, clothes, coals and light, sundries. The expenditure, x, in each budget would thus be expressed as the sum of five variables, a, b, c, d, e, and the mean of the 100 different x's would equal the sum of the means of the a's, the b's, the c's, the d's, and the e's.

(2) *Sets of observations are made which differ in locality or time or some other respect. To find the resultant mean.*

Let l observations of the variable x refer, say, to one date,
„ m „ „ „ „ „ a second „
„ n „ „ „ „ „ a third „

and so on, and let the means of these successive groups of observations be $\bar{x}_l$, $\bar{x}_m$, $\bar{x}_n$, . . . , so that we may write

$$\bar{x}_l=\Sigma x_l/l,\ \bar{x}_m=\Sigma x_m/m,\ \bar{x}_n=\Sigma x_n/n,\ \ldots$$

If then $\bar{x}$ be the resultant mean, we have

$$\bar{x}=\frac{\Sigma x_l+\Sigma x_m+\ldots}{l+m+\ldots}=\frac{l\bar{x}_l+m\bar{x}_m+\ldots}{l+m+\ldots}$$

Example.—If the school children in the different schools of a county are weighed, l children in one school, m in another, n in another, and so on, giving mean weights $\bar{x}_l$, $\bar{x}_m$, $\bar{x}_n$ . . . , the resultant mean weight for the children in all the schools combined is then given by the above expression.

5. **Mean and Standard Deviation of a Distribution of Variables.** Let x_1, x_2, x_3 . . . x_n denote the deviations of each value, or group mid-value, of the observed organ or character when measured from some fixed value, and let f_1, f_2, f_3 . . . f_n denote the observed frequencies of these respective deviations.

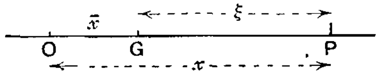
The arithmetic mean of the variables is thus given by

$$\bar{x}=(f_1x_1+f_2x_2+\ .\ .\ .\ +f_nx_n)/(f_1+f_2+\ .\ .\ .\ +f_n),$$

referred to the fixed value as origin.

We may conveniently represent the deviations x_1, x_2, x_3 . . . by lengths measured from an arbitrary origin 0 along a straight line, in which case the point 0 defines the position of the fixed value from which the variables are measured.

Let P mark the position corresponding to a typical variable and let G mark the position corresponding to the mean, $\bar{x}$. Thus OP$=x$, OG$=\bar{x}$, and if we denote the distance of P from G by ξ, we have



$$x=\bar{x}+\xi.$$

Hence

$$\begin{aligned}\bar{x}&=(f_1x_1+f_2x_2+\ .\ .\ .\ +f_nx_n)/(f_1+f_2+\ .\ .\ .\ +f_n)\\&=[f_1(\bar{x}+\xi_1)+f_2(\bar{x}+\xi_2)+\ .\ .\ .\ +f_n(\bar{x}+\xi_n)]/(f_1+f_2+\ .\ .\ .\ +f_n)\\&=[\bar{x}(f_1+f_2+\ .\ .\ .\ +f_n)+(f_1\xi_1+f_2\xi_2+\ .\ .\ .\ +f_n\xi_n)]/f_1+f_2+\ .\ .\ .\ +f_n)\\&=\bar{x}+(f_1\xi_1+f_2\xi_2+\ .\ .\ .\ +f_n\xi_n)/(f_1+f_2+\ .\ .\ .\ +f_n).\end{aligned}$$

Therefore $(f_1\xi_1+f_2\xi_2+\ .\ .\ .\ +f_n\xi_n)=0$ . . . . . (1)

The expression $(f_1x_1+f_2x_2+\ .\ .\ .\ +f_nx_n)$ is called the *first moment of the distribution referred to* 0 *as origin.* We conclude that when the distribution is referred to G as origin, *i.e. when deviations are measured from the mean of the distribution, the first moment vanishes.*

FREQUENCY DISTRIBUTION TABLE.

(1) Deviations of Variables from some fixed value.	(2) Frequency of Deviations.	(3) Product of Nos. in Col. (1) and Col. (2).	(4) Product of Nos. in Col. (1) and Col. (3).
x_1	f_1	f_1x_1	$f_1x_1^2$
x_2	f_2	f_2x_2	$f_2x_2^2$
x_3	f_3	f_3x_3	$f_3x_3^2$
..	..	..	..
..	..	..	..
x_n	f_n	f_nx_n	$f_nx_n^2$
..	N	N'_1	N'_2

In the notation of the above table, where the dashes are omitted in N_1, N_2 when the mean is origin, we have

$$\bar{x}=N'_1/N \text{ and } N_1=0.$$

Again, the root-mean-square deviation, s, measured from the arbitrary origin 0, is given by

$$s^2=(f_1x^2_1+f_2x^2_2+\ \ldots\ +f_nx_n^2)/(f_1+f_2+\ \ldots\ +f_n)$$
$$=\mathrm{N}'_2/\mathrm{N},$$

and N'_2 is called the *second moment of the distribution referred to* 0 as origin.

Substituting as before we have

$$s^2=[f_1(\bar{x}+\xi_1)^2+\ \ldots\ +f_n(\bar{x}+\xi_n)^2]/(f_1+\ \ldots\ +f_n)$$
$$=\frac{\bar{x}^2(f_1+\ \ldots\ +f_n)+2\bar{x}(f_1\xi_1+\ \ldots\ +f_n\xi_n)+(f_1\xi^2_1+\ \ldots\ +f_n\xi_n^2)}{(f_1+\ \ldots\ +f_n)}$$
$$=\bar{x}^2+(f_1\xi^2_1+\ \ldots\ +f_n\xi_n^2)/(f_1+\ \ldots\ +f_n),$$

since $\qquad f_1\xi_1+\ \ldots\ +f_n\xi_n=0.$

Hence $\qquad \mathbf{s}^2=\bar{\mathbf{x}}^2+\sigma^2, \quad . \quad . \quad . \quad (2)$

where σ is the root-mean-square deviation measured from G as origin, or *the standard deviation* as it is called.

From this result it is clear that σ is always less than s, or *the root-mean-square deviation is least when measured from the arithmetic mean.*

Generally, if we write

$$\nu_k'=(f_1x_1^k+\ \ldots\ +f_nx_n^k)/(f_1+\ \ldots\ +f_n),$$
$$\nu_k=(f_1\xi_1^k+\ \ldots\ f_n\xi_n^k)/(f_1+\ \ldots\ +f_n),$$

where $\Sigma(fx^k)$ and $\Sigma(f\xi^k)$ may be called the kth moments referred to 0 and to the mean as origins respectively, so that $\nu_1=0$, $\nu_2=\sigma^2$, $\nu'_2=s^2$, we have

$$\nu'_k=[f_1(\xi_1+\bar{x})^k+\ \ldots\ +f_n(\xi_n+\bar{x})^k]/(f_1+\ \ldots\ +f_n)$$
$$=\frac{\left[(f_1\xi_1^k+\ \ldots\ +f_n\xi_n^k)+k(f_1\xi_1^{k-1}+\ \ldots\ +f_n\xi_n^{k-1})\bar{x}+\frac{k(k-1)}{1\cdot2}(f_1\xi_1^{k-2}+\ldots+f_n\xi_n^{k-2})\bar{x}^2+\ldots+(f_1+\ldots+f_n)\bar{x}^k\right]}{(f_1+\ \ldots\ +f_n)}$$
$$=\nu_k+k\nu_{k-1}\bar{x}+\frac{k(k-1)}{1\cdot2}\nu_{k-2}\cdot\bar{x}^2+\ \ldots\ +\bar{x}^k.$$

For example, when $k=2$, since $\nu_0=1$ and $\nu_1=0$,

$$\nu_2=\nu'_2-\bar{\mathbf{x}}^2 \quad . \quad . \quad . \quad (2)\text{ bis}$$

Again, when $k=3$, $\qquad \nu_3=\nu'_3-3\nu_2\bar{\mathbf{x}}-\bar{\mathbf{x}}^3 \quad . \quad . \quad . \quad (3)$

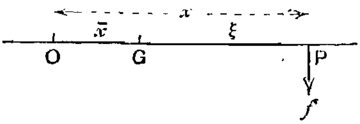
and, when $k=4$, $\qquad \nu_4=\nu'_4-4\nu_3\bar{\mathbf{x}}-6\nu_2\bar{\mathbf{x}}^2-\bar{\mathbf{x}}^4 \quad . \quad . \quad (4)$

There are interesting statical analogues to the above results concerning the mean and standard deviation.

Let us imagine a set of weights, f_1, f_2, f_3 . . . suspended at P_1, P_2, P_3 . . . from a straight horizontal bar, and let the distance of any typical weight f from some arbitrary origin 0 on the bar be x. Then the first moment,

$$f_1x_1+f_2x_2+ \ . \ . \ . \ +f_nx_n$$

(where some of the x's may be negative corresponding to weights suspended to the left of 0) measures the total turning effect of all the given weights about 0, and if we further imagine all these weights replaced by a single weight equal to their sum (f_1+f_2+ . . . $+f_n$), then, in order to produce the same turning effect, it would have to be placed at a point G, the distance of which from 0 is given by



$$\bar{x}(f_1+f_2+ \ . \ . \ . \ +f_n)=(f_1x_1+f_2x_2+ \ . \ . \ . \ +f_nx_n).$$

Thus $\bar{x}=(f_1x_1+f_2x_2+ \ . \ . \ . \ +f_nx_n)/(f_1+f_2+ \ . \ . \ . \ +f_n)$,

and, statically, this defines the position of the centre of gravity of the given weights, f_1, f_2, . . . f_n, relative to 0.

As before, $\bar{x}=\Sigma f(\bar{x}+\xi)/\Sigma f$

$=\bar{x}+\Sigma(f\xi)/\Sigma f$;

hence $f_1\xi_1+f_2\xi_2+ \ . \ . \ . \ +f_n\xi_n=0$,

and, statically, this means that the turning effect of f_1, f_2 . . . f_n about G is zero, in other words, the bar would balance freely about G.

Again, the second moment,

$$f_1x^2{}_1+f_2x^2{}_2+ \ . \ . \ . \ +f_nx_n{}^2,$$

measures the moment of inertia of the weights f_1, f_2 . . . f_n about 0, and, if we imagine these different weights replaced by a single weight (f_1+f_2+ . . . $+f_n$) as before, the moment of inertia will be unaltered if the latter be located at a distance s from 0, where

$$(f_1+f_2+ \ . \ . \ . \ +f_n)s^2=(f_1x^2{}_1+f_2x^2{}_2+ \ . \ . \ . \ +f_nx_n{}^2);$$

therefore $s^2=(f_1x^2{}_1+ \ . \ . \ . \ +f_nx_n{}^2)/(f_1+ \ . \ . \ . \ +f_n)$

$=\Sigma f(\bar{x}+\xi)^2/\Sigma f$

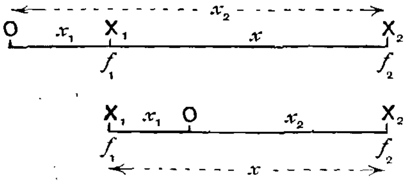
$=\bar{x}^2+\sigma^2$,

as before, and the interpretation of this is that the square of the radius of gyration of the system of weights about 0 equals the square of the radius of gyration about G, the centre of gravity of the system, together with the square of the distance of G from 0. Also, s is clearly least when it is measured from G.

6. The Mean Deviation a Minimum when measured from the Median. Consider first the case when only two different values of the variable are observed, X_1, X_2, and let their deviations from an arbitrary value, 0, chosen as origin, be respectively x_1, x_2.

If f_1, f_2 be the observed frequencies of these values, the sum of their deviations from 0 is

$$(f_1x_1+f_2x_2),$$



which is clearly less when the value 0 lies between X_1, X_2 than when it is smaller or greater than both of them.

Choosing 0, therefore, between X_1, X_2, if f_1 be the **greater** frequency we write the deviation sum

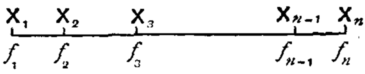
$$=f_1x_1+f_2(x-x_1)$$
$$=f_2x+(f_1-f_2)x_1,$$

where x is the deviation of either of the values X_1, X_2 from the other, and (f_1-f_2) is positive since $f_1>f_2$.

Now this is evidently least when $(f_1-f_2)x_1$ vanishes, *i.e.* when (1) $x_1=0$, in which case 0 coincides with X_1, the more frequent of the two variables, or, when (2) $f_1=f_2$, and in this case, when the two observed values occur equally often, the deviation sum is constant for *any* origin between X_1 and X_2.

When several different values of the variable are observed, they may be arranged in order of magnitude, X_1, X_2, X_3 . . . X_n, from the least to the greatest, with frequencies f_1, f_2, f_3 . . . f_n.

If $f_1>f_n$ we pair off f_n of the X_n's with f_n of the X_1's ; the deviation sum for this pair is least and remains constant when measured from any origin between X_1 and X_n. We next pair off some or all of the X_1's which remain against



an equal number of X_{n-1}'s and the deviation sum for this pair is least and remains constant when measured from any origin between X_1 and X_{n-1}. If some X_1's still remain, we pair them off so far as we can against an equal number of X_{n-2}'s but, if it be X_{n-1}'s that remain, we pair them off against an equal number of X_2's.

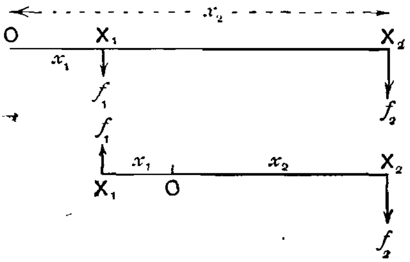
This process can evidently be continued until ultimately we reach the origin from which the mean deviation of the whole distribution is a minimum, for if any X be left unpaired the origin will coincide with that X. Otherwise, the deviation is least when

measured from *any* value between the last two X's paired off together, and within that range it is constant.

Since, by definition, the median is the value of the variable half-way along the series of given observations, ranged in order of their magnitude and assigning each its due weight or frequency, it is clearly such that a balance can be effected by pairing off the values on either side of it against one another in the manner explained above ; it therefore follows that the mean deviation of a frequency distribution is a minimum when the deviations are measured from the median.

The statical analogy to the median also is worth noting. With the same notation as before, the moment or turning effect of two forces, f_1, f_2, about 0 is

$$f_1x_1+f_2x_2.$$



But in this case, if 0 be taken at some point in between X_1 and X_2, since the mean deviation sums the separate deviations without regard to sign, we must imagine f_1 reversed so as to produce a turning effect in the same direction as before. The moment will then be still $(f_1x_1+f_2x_2)$, and it is less when 0 occupies such a position than when it is on X_1X_2 produced in either direction.

Taking 0, therefore, somewhere in between X_1 and X_2, the moment may be written

$$=f_2(x_1+x_2)+x_1(f_1-f_2) ;$$

and, if $f_1>f_2$, this is least when x_1 vanishes, that is, when 0 coincides with X_1, but if $f_1=f_2$, the two forces constitute a *couple*, and the moment is the same whatever position 0 occupies between X_1 and X_2.

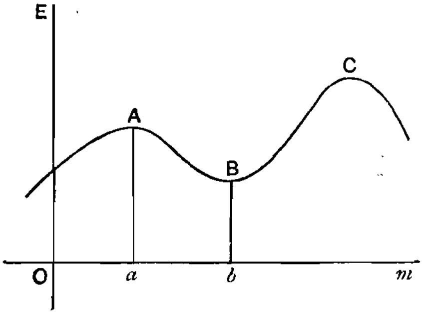
7. **The Method of Least Squares.** To the student who is unacquainted with the differential calculus, the following descriptive argument, the basis of the principle of least squares, for determining the values of m and c which make

$$(mx_1+c-y_1)^2+(mx_2+c-y_2)^2+ \ . \ . \ . \ +(mx_n+c-y_n)^2 \quad . \ . \ . \ (1)$$

a minimum, may prove instructive.

Let us call the above expression E and let us suppose that different values are given to m while c remains unchanged ; in that case E

will vary with m, and we might imagine the different values obtained for E plotted against the corresponding values of m giving a curve of some type. Such a curve may rise and fall in wave-like fashion as in the figure, resulting in maximum points like A and C, and minimum points like B, where we define a maximum point to be such that, as we move away from it along the curve, whether to left or right, the size of the ordinate (and therefore the value of E) decreases; likewise, a minimum point is such that, as we move away from it, the ordinate (and therefore also E) increases. In the neighbourhood of such points it is clear that the size of the ordinate, such as Aa or Bb, changes so slowly as to be practically stationary.



Suppose then that m and $(m+\mu)$, μ being very small, are two values of m respectively at and near a minimum position on the curve, *i.e.* a position like B corresponding to a minimum value for E. Since E *near* such a point does not differ appreciably from E *at* such a point, we may practically equate the two expressions obtained for E by substituting $(m+\mu)$ and m respectively for m in (1), thus

$$(\overline{m+\mu}x_1+c-y_1)^2+(\overline{m+\mu}x_2+c-y_2)^2+\ \dots$$
$$=(mx_1+c-y_1)^2+(mx_2+c-y_2)^2+\ \dots$$

$$(\overline{mx_1+c-y_1+\mu x_1})^2+(\overline{mx_2+c-y_2+\mu x_2})^2+\ \dots$$
$$=(mx_1+c-y_1)^2+(mx_2+c-y_2)^2+\ \dots$$

$$[(mx_1+c-y_1)^2+2\mu x_1(mx_1+c-y_1)+\mu^2x^2_1]+\ \dots$$
$$=(mx_1+c-y_1)^2+\ \dots$$

Thus $[2x_1(mx_1+c-y_1)+\mu x^2_1]+\ \dots\ =0.$

Now, the smaller we take μ, the nearer to the truth does this result become. Hence, by making μ tend to zero, we are led to the strictly true relation

$$x_1(mx_1+c-y_1)+\ \dots\ =0.$$

This is one of the equations in the text. To obtain the second, we keep m constant and vary c.

Suppose c and $(c+\gamma)$ are two values of c at and near a minimum

position on the curve; then, equating the two corresponding values of E, we have as before

$$(mx_1+\overline{c+\gamma}-y_1)^2+\ \ldots\ =(mx_1+c-y_1)^2+\ \ldots$$
$$(\overline{mx_1+c-y_1}+\gamma)^2+\ \ldots\ =(mx_1+c-y_1)^2+\ \ldots$$
$$[(mx_1+c-y_1)^2+2\gamma(mx_1+c-y_1)+\gamma^2]+\ \ldots\ =(mx_1+c-y_1)^2+\ \ldots$$

Thus $$[2(mx_1+c-y_1)+\gamma]+\ \ldots\ =0.$$

and, proceeding to the limit when γ tends to zero, we reach the other equation in the text, namely,

$$(mx_1+c-y_1)+\ \ldots\ =0.$$

[The Method of Least Squares came first into prominence in Astronomy in connection with the determination of the best value to take when a number of observations, apparently equally reliable, give results not quite in agreement. If, for instance, x be the true value of some variable, and if x_1, x_2, x_3 . . . x_n be the results of n observations, the method of least squares assumes x to be given by making

$$y=(x-x_1)^2+(x-x_2)^2+\ \ldots\ +(x-x_n)^2$$

a minimum.

Now $\dfrac{dy}{dx}=2(x-x_1)+2(x-x_2)+\ \ldots\ +2(x-x_n)$, and this vanishes

when $$(x-x_1)+(x-x_2)+\ \ldots\ +(x-x_n)=0,$$

i.e. $$x=(x_1+x_2+\ \ldots\ +x_n)/n,$$

so that in this case we are led to the ordinary arithmetic mean of the n observations as the best value.

The method was used by Gauss as early as 1795.]

8. To prove $\int_{-\infty}^{+\infty} e^{-x^2}dx=\sqrt{\pi}$.

Let $$I=\int_{-\infty}^{+\infty} e^{-x^2}dx\,;$$

thus, also, $$I=\int_{-\infty}^{+\infty} e^{-y^2}dy\,;$$

therefore, $$I^2=\int_{-\infty}^{+\infty} e^{-x^2}dx\int_{-\infty}^{+\infty} e^{-y^2}dy$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)}dxdy$$

$$=\int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-r^2}rdrd\theta$$

(by changing to polar co-ordinates)

$$=\int_0^\infty e^{-r^2} r\,dr \int_0^{2\pi} d\theta$$

$$=\left[-\frac{e^{-r^2}}{2}\right]_0^\infty \left[\theta\right]_0^{2\pi}$$

$$=(\tfrac{1}{2})(2\pi).$$

Hence $$I=\int_{-\infty}^{+\infty} e^{-x^2}dx=\sqrt{\pi}.$$

9. To prove :—

(1) $\Gamma(n+1)=n\Gamma(n)$. (2) $B(m, n)=\dfrac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$.

(1) $$\Gamma(n+1)=\int_0^\infty x^n e^{-x}dx$$

$$=-\int_{x=0}^\infty x^n d(e^{-x})$$

$$=\left[-x^n e^{-x}\right]_{x=0}^\infty + n\int_0^\infty x^{n-1}e^{-x}dx$$

$$=n\Gamma(n),$$

because the expression in square brackets vanishes at both limits.

(2) $$\Gamma(m)\Gamma(n)=\int_0^\infty e^{-\xi}\xi^{m-1}d\xi \int_0^\infty e^{-\eta}\eta^{n-1}d\eta$$

$$=\int_0^\infty e^{-x^2}x^{2m-2}2x\,dx\int_0^\infty e^{-y^2}y^{2n-2}2y\,dy,$$

where $x^2=\xi$, $y^2=\eta$.

Hence $$\Gamma(m)\Gamma(n)=4\int_0^\infty\int_0^\infty e^{-(x^2+y^2)}x^{2m-1}y^{2n-1}dx\,dy$$

$$=\int_{r=0}^\infty\int_{\theta=0}^{2\pi} e^{-r^2}r^{2m+2n-2}\cos^{2m-1}\theta\sin^{2n-1}\theta\, r\,dr\,d\theta$$

(by changing to polar co-ordinates).

Thus $$\Gamma(m)\Gamma(n)=\int_0^\infty e^{-r^2}r^{2m+2n-1}dr\int_0^{2\pi}\cos^{2m-1}\theta\sin^{2n-1}\theta\,d\theta$$

$$=\left[\frac{1}{2}\int_0^\infty e^{-p}p^{m+n-1}dp\right].\left[\frac{4}{2}\int_0^1 k^{n-1}(1-k)^{m-1}dk\right],$$

where $p=r^2$ and $k=\sin^2\theta$;

therefore, $$\Gamma(m)\Gamma(n)=\Gamma(m+n)B(n, m)$$

$$=\Gamma(m+n)B(m, n)$$

by symmetry.

10. Elementary Method of Testing the Probability Integral Table. The reader may find more satisfaction in using the probability integral table if he tests for himself one or two of its results by means of squared paper or in some other way.

We have seen that the probability of an error between 0 and $\sigma\xi$ is given by the expression

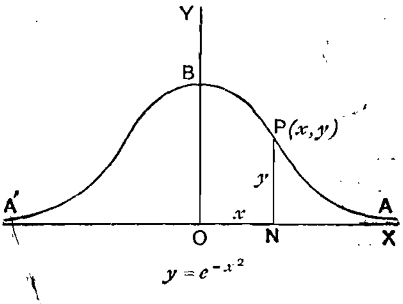
$$\frac{1}{\sqrt{2\pi}}\int_0^{\xi} e^{-\frac{1}{2}\xi^2}d\xi.$$

Put $\xi=\sqrt{2}x$, and this becomes

$$\frac{1}{\sqrt{\pi}}\int_0^{x} e^{-x^2}dx=\int_0^{x} e^{-x^2}dx \Big/ \int_{-\infty}^{+\infty} e^{-x^2}dx, \text{ by Note (8)}$$

$$=\text{area OBPN/area A}'\text{BA, in the figure.}$$

Now the graph of $y=e^{-x^2}$ is drawn in fig. (40) of the text, and it is possible therefore to get an approximation to the above result for any value of x by counting the number of small squares in that figure enclosed by the areas corresponding to OBPN and A′BA respectively. Each complete small square may be reckoned as 1, and each portion of a square may be reckoned as 1 if it exceeds half a square and as zero if it is less than half a square.



This gives, for example,

$$\frac{1}{\sqrt{\pi}}\int_0^{0\cdot25} e^{-x^2}dx=98/707=0\cdot139,$$

whereas the tables give 0·138.

For a value like $x=0\cdot71$, count the squares in the usual way between curve, axes, and ordinate $x=0\cdot70$; then add to the result one-fifth of the number of squares in the small slice of area between curve, axis, and ordinates $x=0\cdot70$ and $x=0\cdot75$. We get

$$\frac{1}{\sqrt{\pi}}\int_0^{0\cdot71} e^{-x^2}dx=240/707=0\cdot339$$

as compared with 0·342 from the tables.

These results are not unsatisfactory considering the rough nature of the method followed to obtain them.

11. **The Law of Frequency in the case of two Correlated Variables with certain Deductions therefrom**—[based on Professor Karl Pearson's memoir, *Regression, Heredity and Panmixia* (*Phil. Trans.*, vol. 187A, pp. 253-318)].

Consider two variables whose deviations, x and y, from their respective means are due to a number of *independent* causes, the deviations in which from their means can be quantitatively denoted by $\epsilon_1, \epsilon_2, \ldots \epsilon_m$.

We assume that each ϵ deviation is so small compared to the mean value from which it is measured that x and y can be sensibly expressed as linear functions, thus

$$x = a_1\epsilon_1 + a_2\epsilon_2 + \ldots + a_m\epsilon_m \quad . \quad . \quad . \quad (1)$$

$$y = b_1\epsilon_1 + b_2\epsilon_2 + \ldots + b_m\epsilon_m \quad . \quad . \quad . \quad (2)$$

(Some of the a's and b's may be zero, and if x only involved, say, $\epsilon_1, \epsilon_2 \ldots \epsilon_k$, and y only involved $\epsilon_{k+1} \ldots \epsilon_m$, then it would be natural to expect no correlation between x and y.)

We further assume that each ϵ varies according to the normal law with S.D. σ with appropriate suffix.

Equations (1) and (2) show that the same x and y may arise in a multitude of different ways obtained by varying the ϵ's so that their weighted sums (the a's and b's being the weights) remain unaltered. The probability that the particular deviations lying between

$$\epsilon_1 \smile (\epsilon_1 + \delta\epsilon_1), \ \epsilon_2 \smile (\epsilon_2 + \delta\epsilon_2), \ \ldots \ \epsilon_m \smile (\epsilon_m + \delta\epsilon_m)$$

shall concur, since they are all independent, is

$$z = \left(\frac{\delta\epsilon_1}{\sigma_1\sqrt{2\pi}} e^{-\epsilon_1^2/2\sigma_1^2}\right) \ldots \left(\frac{\delta\epsilon_m}{\sigma_m\sqrt{2\pi}} e^{-\epsilon_m^2/2\sigma_m^2}\right).$$

But, writing

$$a_3\epsilon_3 + \ldots + a_m\epsilon_m = \alpha, \quad b_3\epsilon_3 + \ldots + b_m\epsilon_m = \beta,$$

$\therefore$ equations (1) and (2) become

$$a_1\epsilon_1 + a_2\epsilon_2 + (\alpha - x) = 0$$

$$b_1\epsilon_1 + b_2\epsilon_2 + (\beta - y) = 0.$$

Therefore $$\frac{\epsilon_1}{a_2(\beta - y) - b_2(\alpha - x)} = \frac{\epsilon_2}{b_1(\alpha - x) - a_1(\beta - y)} = \frac{1}{a_1b_2 - a_2b_1}.$$

And, for any function ν,

$$\iint \nu\, dx dy = \iint \nu \left(\frac{\delta x}{\delta\epsilon_1}\frac{\delta y}{\delta\epsilon_2} - \frac{\delta x}{\delta\epsilon_2}\frac{\delta y}{\delta\epsilon_1}\right) d\epsilon_1 d\epsilon_2$$

$$= (a_1b_2 - a_2b_1) \iint \nu\, d\epsilon_1 d\epsilon_2.$$

Hence

$$z = \frac{\delta x \delta y}{a_1 b_2 - a_2 b_1} \cdot \frac{e^{-\left(\frac{\epsilon_3^2}{2\sigma_3^2} + \cdots + \frac{\epsilon_m^2}{2\sigma_m^2}\right)}}{(\sigma_1 \ . \ . \ . \ \sigma_m)(2\pi)^{m/2}} \cdot$$

$$e^{-\frac{(b_2 x - a_2 y + a_2\beta - b_2\alpha)^2}{2\sigma_1^2(a_1 b_2 - a_2 b_1)^2} - \frac{(a_1 y - b_1 x + b_1\alpha - a_1\beta)^2}{2\sigma_2^2(a_1 b_2 - a_2 b_1)^2}} \cdot \delta\epsilon_3 \ . \ . \ . \ \delta\epsilon_m.$$

The *total* probability for deviations between $x_{\smile}(x+\delta x)$ and $y_{\smile}(y+\delta y)$ is obtained by integrating z between limits $-\infty$ and $+\infty$ for all the ϵ's from ϵ_3 to ϵ_m, and it is not very difficult to see that this will ultimately lead to an expression of the form

$$\mathbf{C} \cdot \delta \mathbf{x} \delta \mathbf{y} \cdot \mathbf{e}^{-(\mathbf{ax^2} + \mathbf{2hxy} + \mathbf{by^2})}.$$

This is the required law of frequency.

To find the meanings of the constants a, b, h. The total probability for a deviation between $x_{\smile}(x+\delta x)$ associated with *any* deviation y is

$$= C\delta x \int_{-\infty}^{+\infty} e^{-(ax^2 + 2hxy + by^2)} dy$$

$$= C\delta x \int_{-\infty}^{+\infty} e^{-b\left\{\left(y + \frac{hx}{b}\right)^2 + x^2\left(\frac{a}{b} - \frac{h^2}{b^2}\right)\right\}} dy$$

$$= C\delta x \cdot e^{-x^2\left(\frac{ab-h^2}{b}\right)} \int_{-\infty}^{+\infty} e^{-b\left(y + \frac{hx}{b}\right)^2} dy$$

$$= C\sqrt{\pi/b}\,\delta x e^{-x^2(ab-h^2)/b}.$$

But if x be subject to the normal law, the probability for a deviation between $x_{\smile}(x+\delta x)$ is

$$\frac{\delta x}{\sqrt{2\pi} \cdot \sigma_x} e^{-x^2/2\sigma_x^2},$$

where σ_x is the S.D. of x independent of y.

Comparing these two results, we have

$$\mathbf{1/2\sigma_x^2 = (ab - h^2)/b = a(1 - r^2)},$$

if $$r = -h/\sqrt{ab}.$$

Similarly, $$\mathbf{1/2\sigma_y^2 = (ab - h^2)/a = b(1 - r^2)},$$

so that $$\mathbf{h = -r\sqrt{ab} = -r/2\sigma_x\sigma_y(1 - r^2)}.$$

Again, we may integrate z for all values of x and y, and so get the total frequency, N, of the (x, y) pair.

Thus, $$\mathrm{N} = C\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(ax^2 + 2hxy + by^2)} dx dy$$

$$= C\sqrt{\pi/b}\int_{-\infty}^{+\infty} e^{-x^2(ab-h^2)/b} dx$$

$$= C\sqrt{\pi/b}\sqrt{\pi}\sqrt{[b/(ab - h^2)]}.$$

Hence
$$C=\frac{N}{\pi}\sqrt{(ab-h^2)}$$
$$=\frac{N}{\pi}\sqrt{[ab(1-r^2)]}$$
$$=\frac{N}{2\pi\sigma_x\sigma_y\sqrt{(1-r^2)}}.$$

Thus
$$z=Ce^{-\frac{1}{2(1-r^2)}\left[\frac{x^2}{\sigma_x^2}-\frac{2rxy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\right]}\delta x\delta y,$$

where C has the above value.

It still remains to interpret r and to see that it is really the coefficient of correlation as defined in Chapter X. For this purpose let us suppose we have observed n pairs of associated x's and y's, namely

$$(x_1y_1),\ (x_2y_2)\ .\ .\ .\ (x_ny_n).$$

The probability for such a concurrence, taken along with a given value for r and assuming the observations independent, is proportional to

$$\frac{1}{\sqrt{(1-r^2)}}\cdot e^{-\frac{1}{2(1-r^2)}\left[\frac{x_1^2}{\sigma_x^2}-\frac{2rx_1y_1}{\sigma_x\sigma_y}+\frac{y_1^2}{\sigma_y^2}\right]}\times\ldots\times\frac{1}{\sqrt{(1-r^2)}}e^{-\frac{1}{2(1-r^2)}\left[\frac{x_n^2}{\sigma_x^2}-\frac{2rx_ny_n}{\sigma_x\sigma_y}+\frac{y_n^2}{\sigma_y^2}\right]}$$
$$=\frac{1}{(1-r^2)^{n/2}}e^{-\frac{1}{2(1-r^2)}\left[\frac{\Sigma x^2}{\sigma_x^2}-\frac{2r\Sigma xy}{\sigma_x\sigma_y}+\frac{\Sigma y^2}{\sigma_y^2}\right]}$$
$$=(1-r^2)^{-n/2}e^{-\frac{1}{2(1-r^2)}[2n-2rn\kappa]},$$

where $\kappa=\Sigma xy/n\sigma_x\sigma_y$

$$=e^{-\frac{n}{2}\log(1-r^2)-\frac{n}{1-r^2}(1-\kappa r)}.$$

Now the probability of this particular distribution is greatest when

$$\tfrac{1}{2}\log(1-r^2)+\frac{1-\kappa r}{1-r^2}$$

is least, and, differentiating with respect to r, this leads to

$$\tfrac{1}{2}\frac{-2r}{1-r^2}+\frac{(1-r^2)(-\kappa)+2r(1-\kappa r)}{(1-r^2)^2}=0,$$

i.e. $\quad -r(1-r^2)-\kappa(1-r^2)+2r(1-\kappa r)=0,$

i.e. $\quad -r+r^3-\kappa+\kappa r^2+2r-2\kappa r^2=0,$

i.e. $\quad (r-\kappa)(1+r^2)=0.$

It is not difficult to show that $r=\kappa$ gives a minimum; hence the required probability is a maximum and we get the best value for the coefficient by taking

$$\mathbf{r}=\kappa=\Sigma\mathbf{xy}/\mathbf{n}\sigma_x\sigma_y.$$

CERTAIN CURRENT SOURCES OF SOCIAL STATISTICS

Any one who is anxious to get reliable figures bearing upon some social matter is somewhat at a pause unless he is thoroughly conversant with all the statistical ramifications of Government authorities, local and national, of trade unions, friendly societies, and hosts of other bodies of a public or semi-public character.

While recognizing the lavish outpouring of statistics of all kinds upon a multitude of diverse topics every year, and appreciating the immense care and patience shown by those who are responsible for their collection and preparation, one cannot but deplore the lack of any co-ordinating principle in general between one body and another either in deciding what statistics shall be collected, by whom and when they shall be collected, or how afterwards they shall be tabulated and presented to the public. Too often a narrow-minded jealousy prevents one authority from consulting with another, and such co-operation as does exist is due largely to the efforts of able and enlightened individuals. The result is that a vast amount of labour and expense goes waste and the loss to the public is incalculable, but the public do not care, and they do not care because they do not know.

At present, to quote from an influential petition on the subject recently presented to His Majesty's Government, 'It is almost universally the case that any serious investigation is reduced to roughly approximate estimates in relation to some factor which is essential for its result. . . . It is not too much to say that there is hardly any reform, financial, social, or commercial, for which adequate information can be provided with our present machinery.' But this state of things would be partly remedied by adequate control such as might be secured by the establishment of a central statistical office with a minister in charge who should be responsible for unification so far as possible in the collection, tabulation, and issue of all public statistics.

It is scarcely possible for a single private individual to make a quantitative investigation of any social question on a large enough scale to produce results of real value; conspicuous instances like Booth and Rowntree might seem to be exceptions to this rule, but even they had a number of workers acting under their direction, without whose aid their task would have seemed almost hopeless.

For such statistics as we have we are therefore dependent upon Government departments, local authorities, public officials, trade associations representing employers or labour, public companies, and so on. The reader who wishes to get some idea of the extent and the limitations of official British statistics is referred to the admirable introductory chapters of Bowley's *Elements of Statistics.* Here we cannot do more than mention a very few of the most important sources whence such statistics are derived.

The most voluminous of all our records is probably the *Census of the Population* which is taken every ten years. Its scope is but faintly realized by enumerating the chief subjects on which the Registrar-General asked information from each householder in 1911, namely :

(1) Numbers and Geographical Distribution of the Population.
(2) Nationality and Birth-place.
(3) Numbers at Different Ages, Male and Female.
(4) Numbers Single, Married, and Widowed.
(5) Sizes of Families, including Children Dead.
(6) Numbers engaged in different Professions and Occupations.
(7) Numbers Blind, Deaf, Dumb, not in their Right Mind.
(8) Numbers occupying Dwellings of Different Sizes as measured by the Number of Rooms.

This may seem an ambitious scheme when it is stated that the mere enumeration of the people was successfully opposed less than two hundred years ago as 'subversive of the last remains of English liberty and likely to result in some public misfortune or an epidemical disorder,' and the first census was only taken in 1801. [See Article in the *Encyclopaedia Britannica* on the subject.]

The results of each census are published in bulky volumes as soon as they can be reduced and tabulated, a process which, of course, takes a considerable time even for an army of workers with calculating machines and every modern device to facilitate their progress. It is to be regretted that more is not done to advertise so valuable a record of work by publication in a cheap and attractive form of a summary of matters which vitally affect the good of the commonwealth. As it is, the census volumes tend to be purchased only by public authorities and officials who require to use them occasionally as books of reference.

Neglect of the blandishments of advertisement—to be commended in general because such neglect is somehow associated with the presentation of all truth—may be perhaps carried too far in the issue of statistics.

It will be noted that in the periodical census no mention is made of wages though the people are classified as regards occupation, and for information upon this point we must turn to another source. The last general census of wages was taken in 1906, following and improving upon an earlier inquiry twenty years before, but, in connection with an inquiry by the Board of Trade into the cost of living of the working classes, information was collected as to rates of wages in 1912 of workpeople in certain occupations in the building, engineering, and printing trades, these being selected as industries common to most towns, and because the time rates of wages paid in them are largely standardized.

The 1906 *inquiry into earnings and hours of labour*, unlike the decennial census, was conducted on a voluntary basis and was never wholly completed. In brief it set out to discover from employers :—

(1) The Numbers of Working-people Employed in Various Occupations, distinguishing Men, Women, Lads, and Girls.

(2) The Nature of the Work done and the Rates of Wages Paid, distinguishing Time Rates from Piece Rates.

(3) The Hours Worked, distinguishing Under- or Over-time from Normal Time.

The ground actually covered by the inquiry embraces the following trades: Textiles, Clothing, Building and Woodworking, Public Utility Services, Metal, Engineering, and Shipbuilding—in 1906; also Agriculture, and Railway Service—in 1907 ; the reports upon these trades were published separately at different dates between 1909 and 1912, and the following trades were bulked together in one volume, published in 1913—Paper and Printing; Pottery, Brick, Glass, and Chemicals ; Food, Drink, and Tobacco ; and Miscellaneous Trades.

The Cost of Living Inquiry of 1912 was in continuation of a similar inquiry in 1905, which in addition compared conditions in the United Kingdom and certain foreign countries. *It dealt not only with wages but also with rents and retail prices.*

The report states that ‘ particulars as to the rent and accommodation of typical working-class dwellings were obtained from officials of local authorities, surveyors of taxes, house owners and agents, and by house-to-house inquiry.’ Also ‘ returns of the prices most generally paid by working-class customers for a number of specified commodities were obtained in each town by personal inquiry from a number of retailers engaged in working-class trade.’

Since then *Lord Sumner's Committee and a Committee of the*

Agricultural Wages Board have examined *the change in the cost of living between* 1914 *and* 1919, as evidenced by a number of household budgets collected from among urban working-classes and workers in rural districts respectively.

One other highly important inquiry carried out by the Board of Trade deserves notice, namely, the *First Census of Production of the United Kingdom* (1907).

The published report shows :—

(1) The total Net Output in Money Value for each Trade Group in each Industry.

(2) The Number of Persons Employed in each Trade Group (salaried persons and wage-earners exclusive of outworkers).

(3) The Net Output per Person Employed in each Trade Group as deduced from (1) and (2).

(4) The Horse-power of Engines in Mines, Quarries, or Factories Employed in each Trade Group.

It is explained that the term ' net output ' here represents the value of the aggregate output of the factories, etc., from which returns were received in each trade group, after deducting the cost of materials purchased from factories, etc., not included in the group, or supplied by merchants or others not making returns to the Census of Production Office.

Valuable as the results of these inquiries undoubtedly are, they would be of still more value were it only possible satisfactorily to collate the various returns of population, wages, and production. No record of wages was included, for example, in the Census of Production statistics, and it is quite impossible to deduce the number of wage-earners and those dependent upon them in any trade at any given time.

Apart, however, from such special inquiries as we have instanced, and the ten-yearly census of the people, there are other periodical records issued which provide us with valuable information. The Ministry of Labour, until recently a special branch of the Board of Trade, charged with the duty of keeping in touch with labour conditions, issues each month a *Labour Gazette* giving particulars relating to the state of employment in the principal trades in the United Kingdom based on returns from employers, trade unions, and employment exchanges, besides information concerning trade disputes, changes in wages and hours, the course of prices, railway traffic receipts, foreign trade, etc. The Board of Trade also publishes weekly a *Journal and Commercial Gazette* dealing with matters of interest to all who are engaged in commerce or finance ; while a

Monthly Bulletin of Statistics of production, trade, finance, employment, etc., at present issued under the name of the Supreme Economic Council, is an important recent addition to our knowledge of international statistics.

Again the Registrar-General makes a quarterly return and annual summary of births, marriages, and deaths in the different counties of England and Wales, and of births, deaths, and infectious diseases in certain large towns. In each public health area the medical officer reports periodically upon the hygienic condition of the district and the health of the people under his care. The Board of Education is answerable for conditions in the schools, and the Home Office in factories and prisons; they report from time to time. The Ministry of Health similarly issues returns relating to pauperism and to housing, while the Board of Agriculture and Fisheries registers the acreage under crops and the number of live stock in the United Kingdom, and the Commissioners of Customs record the expansion or contraction of foreign trade.

In addition we have the endless accounts and statistics supplied, some voluntarily and some compulsorily, by municipal bodies, public companies, banks, trade associations, co-operative societies, insurance companies, trade unions, etc.

And yet, in spite of all this wealth of statistics, some surprising gaps occur, as we have already seen, in important particulars which cannot be traced. We shall quote only one more instance of such a hiatus—the income-tax returns provide a basis for measuring that part of the national income which is subject to taxation, some idea also can be formed of what the wage-earners receive, but as to the earnings of the portion of the community falling in between these two classes we are entirely ignorant. It is possible that war conditions during the years 1914-19 may have vastly increased the knowledge of the Government as to some matters such as internal resources and inland trade, of which little was known before, but, if so, the public, whom it concerns so closely, have not yet been permitted fully to share in this advantage.

For an excellent summary of labour statistics compiled or collected by the Government the reader is recommended to consult the *Annual Abstract of Labour Statistics of the United Kingdom*, published in the past by the Labour Department of the Board of Trade.

NOTE.—A most useful *Guide to Official Statistics* is now issued by H.M.S.O. Dr. Bowley's *Official Statistics* will repay careful study in conjunction with it.

A NOTE ON TABLES TO AID CALCULATION

The short tables which follow are only inserted as specimens, as it is expected that the reader who wishes to make extensive use of such tables will have access to the fuller ones to which reference is made below.

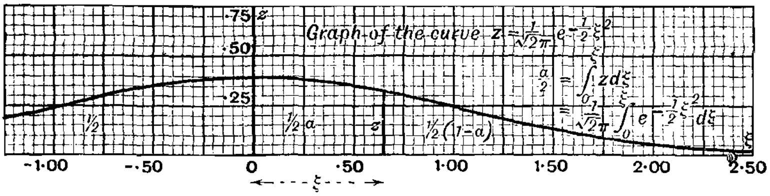


Fig. (55).

Probability Integral Table, giving area of curve $z=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\xi^2}$ in terms of corresponding abscissa, see fig. (55) :—

ξ	$\frac{1}{2}(1+\alpha)$	α	ξ	$\frac{1}{2}(1+\alpha)$	α
·00	·50000	·00000	·76	·77637	·55274
·10	·53983	·07966	·77	·77935	·55870
·20	·57926	·15852	·78	·78230	·56460
·30	·61791	·23582	·79	·78524	·57048
·40	·65542	·31084	·80	·78814	·57628
·45	·67364	·34728	·85	·80234	·60468
·50	·69146	·38292	·90	·81594	·63188
·55	·70884	·41768	·95	·82894	·65788
·60	·72575	·45150	1·00	·84134	·68268
·65	·74215	·48430	1·05	·85314	·70628
·70	·75804	·51608	1·10	·86433	·72866
·71	·76115	·52230	1·50	·93319	·86638
·72	·76424	·52848	2·00	·97725	·95450
·73	·76730	·53460	2·50	·99379	·98758
·74	·77035	·54070	3·00	·99865	·99730
·75	·77337	·54674	3·50	·99977	·99954

Fig. (56), the result of plotting α against ξ, enables us to estimate the probability of an error lying between any two limits.

Table giving P, to test 'goodness of fit,' corresponding to certain values of n' and χ^2 :—

n' ↓	$\chi^2 \rightarrow$ 4	5	6	7	8	9	10	11	12	13	14	15
7	·67668	·54381	·42319	·32085	·23810	·17358	·12465	·08838	·06197	·04304	·02964	·02026
8	·77978	·65996	·53975	·42888	·33259	·25266	·18857	·13862	·10056	·07211	·05118	·03600
9	·85712	·75758	·64723	·53663	·43347	·34230	·26503	·20170	·15120	·11185	·08176	·05914
10	·91141	·83431	·73992	·63712	·53415	·43727	·35048	·27571	·21331	·16261	·12232	·09094
11	·94735	·89118	·81526	·72544	·62884	·53210	·44049	·35752	·28506	·22367	·17299	·13206
12	·96992	·93117	·87336	·79907	·71330	·62189	·53039	·44326	·36264	·29333	·23299	·18250
13	·98344	·95798	·91608	·85761	·78513	·70293	·61596	·52892	·44568	·36904	·30071	·24144
14	·99119	·97519	·94615	·90215	·84360	·77294	·69393	·61082	·52764	·44781	·37384	·30735
15	·99547	·98581	·96649	·93471	·88933	·83105	·76218	·68604	·60630	·52652	·44971	·37815

One of the earliest tables of the probability integral appeared in Kramp's *Analyse des Refractions* (Strasbourg, 1798), where the calculation of $\int_x^\infty e^{-x^2}dx$ was given to eight places from $x=0$ to $x=3$ at intervals of 0·01. Tables more recent and extensive are those due to J. Burgess (*Trans. Roy. Soc. Edin.* 1900) and to W. F. Sheppard (*Biometrika*, vol. ii., pp. 174-190). Of these the latter

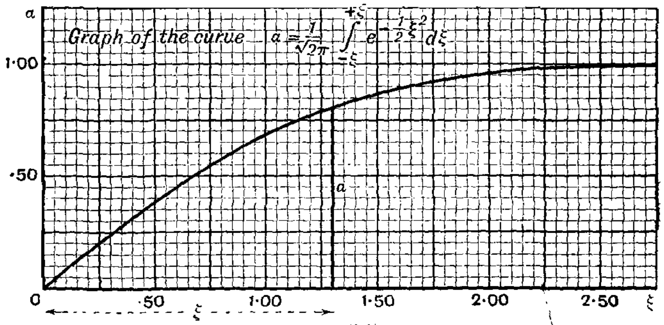


Fig. (56).

is reproduced in the admirable *Tables for Statisticians and Biometricians*, edited by Karl Pearson (Camb. Univ. Press, 1914), and the same volume also contains Palin Elderton's *P Tables* for testing 'goodness of fit' which first appeared in *Biometrika*, vol. i., and Duffell's *Tables of the Logarithms of the Γ Function* from *Biometrika*, vol. vii., besides a large number of other valuable tables.

It should be remarked in connection with the last-named table that the formula $\Gamma(x+1)=x\ \Gamma(x)$ enables us to reduce the calculation of any Γ function to one in which x lies between 1 and 2, by repeated applications of the logarithmic relation, thus

$$\log \Gamma(x+1)=\log x+\log \Gamma(x)$$
$$=\log x+\log (x-1)+\log \Gamma(x-1),$$

and so on. When x is large, however, say greater than 10, the well-known approximate formula

$$\Gamma(x+1)=e^{-x}x^{x+\frac{1}{2}}(2\pi)^{\frac{1}{2}}e^{\frac{1}{12x}}$$

(see, for instance, Whittaker's *Analysis*, § 110) will be found useful, and it may also be written

$$\log \frac{\Gamma(x+1)}{x^x e^{-x}}=0{\cdot}3990899+\frac{0{\cdot}03619121}{x}+\tfrac{1}{2}\log x,$$

a form often convenient.

It may be of service to record here the values of a few constants which frequently recur for speedy reference :

$e=2{\cdot}718\ 2818$	$\pi=3{\cdot}141\ 5926$	$\log_{10} 2=0{\cdot}301\ 0300$
$\frac{1}{e}=0{\cdot}367\ 8794$	$\log_{10}\pi=0{\cdot}497\ 1499$	$\log_{10} 3=0{\cdot}477\ 1213$
$\log_{10} e=0{\cdot}434\ 2945$	$\log_{10}\frac{1}{\sqrt{2\pi}}=\bar{1}{\cdot}600\ 9101$	
$\log_{10}(\log_{10} e)=\bar{1}{\cdot}637\ 7843$		

The statistician who has Pearson's *Tables*, Barlow's *Tables of Squares*, etc., together with a good set of Tables of Logarithms (unless he is so fortunate as to have a mechanical calculator, for instance a Brunsviga, at his disposal) and of Trigonometrical Functions such as Chambers's *Seven-Figure Tables*, may consider himself amply provided for serious research and decidedly better off than his predecessors who prepared the way for him by doing great work with much poorer tools.

MISCELLANEOUS EXAMPLES

[*Selected from London B.Sc.* (*Econ.*) *Pass and Honours papers*]

PART I

(1) Define the genus ' average,' and the principal species of that genus. Adduce concrete cases in which (*a*) the Arithmetic Mean, or (*b*) the Median, is specially appropriate.

(2) Supposing that statistics of rents of working-class dwellings have been collected in a certain district for a series of years, describe some way of forming an index number showing the changes in rents from year to year during the period. Give reasons for the process you adopt, or state any advantages it appears to you to possess.

(3) Measure by whatever method you think most suitable the correlation between the two following series, and show graphically the relationship between the two series.

	Exports per head.	Unemployment Index.		Exports per head.	Unemployment Index.
	£			£	
1884	6·5	8·1	1899	6·5	2·2
5	5·9	9·3	1900	7·1	2·5
6	5·9	10·2	1	6·7	3·3
7	6·1	7·6	2	6·8	4·0
8	6·4	4·6	3	6·9	4·7
9	6·7	2·1	4	7·1	6·0
1890	7·0	2·1	5	7·7	5·0
1	6·5	3·5	6	8·7	3·6
2	6·0	6·3	7	9·7	3·7
3	5·7	7·5	8	8·5	7·8
4	5·6	6·9	9	8·5	7·7
5	5·8	5·8	1910	9·6	4·7
6	6·1	3·4	1	10·0	3·0
7	5·9	3·5	2	10·7	3·2
8	5·8	2·9	3	11·4	2·1

(4) Apply some test by which the figures in the previous table can be used to determine whether unemployment (as there measured) increased or diminished in the 30 years.

(5) Exhibit the difficulty of comparing nations, in respect of power and prosperity, by means of statistics relating to (*a*) the number of population, (*b*) occupations, (*c*) criminality, (*d*) exports and imports.

(6) Draw up, with careful attention to form and detail and showing all sub-totals, a blank table in which could be shown, for the years 1919 to 1923 inclusive, the numbers of students who entered for the Final Examination for B.Sc. (Econ.), distinguishing Internal and External Students, Pass and Honours Candidates, and the results of the examinations (Pass or Fail in the case of Pass Candidates and Honours I, II, III, Pass Degree, Fail in the case of Honours Candidates).

(7) Define the geometric mean, and discuss its use in forming index numbers of prices.

(8) The average prices of wheat and the quantities sold at four markets are given as follows :—

Market.	Average Price per Qr.	Quantity sold, Qrs.
A	27s. 3d.	36,000
B	28s. 8d.	1,000
C	29s. 1d.	16,000
D	27s. 2d.	12,000

Find the mean price for the four markets, weighting each local average with the quantity sold.

Would it be possible for the average price at each of the above markets to rise from one year to the next and yet for the weighted mean price to fall ? If so, under what conditions ?

(9) Illustrate the necessity for standardisation when heterogeneous groups are in question by describing the methods of computing standard birth- or death-rates or family food-consumption.

(10) The following are the Annual Premiums required to secure at death £1000 plus a Guaranteed Reversionary Bonus of £2 per cent. on the sum assured under the Whole Life Policies of a certain Assurance Company :—

Age next Birthday.	Annual Premium.
	£ *s.* *d.*
25	24 12 6
30	27 14 2
35	31 11 8
40	36 7 6
45	42 6 8

Find by any method of interpolation what the premium would be at age 36 next birthday.

(11) Explain why the method of measuring the mortality from any disease by the proportion of deaths from that disease to deaths from all causes is essentially fallacious.

Criticise the following mode of argument in a recent blue-book, containing anthropometric data with respect to school children: The gradation in weight from the poorest group up to the wealthiest is one of the most striking features of the tables. If we take all the children of ages from 5 to 18, we find that the average weight of the boy from a one-roomed tenement is 52·6 lb.; of the boy from a two-roomed tenement, 56·1 lb.; of the boy from a three-roomed tenement, 60·6 lb.; of the boy from a tenement of four rooms or more, 64·3 lb.

(12) Show how to measure the 'trend' and the 'fluctuation' of a series of numbers relating to economic phenomena, such as trade or employment.

(13) Find the average age and the median age of the married men included in the table below, and calculate one measure of dispersion.

Ages.	Married Men.		Widowers.
	Number of Men 000's.	Average Age of Children under 16.	Number of Men 000's.
Under 20	1	·47	..
20—	34	·61	..
25—	99	·97	1
30—	132	1·50	2
35—	139	1·99	3
40—	138	1·98	5
45—	130	1·53	7
50—	104	·95	9
55—	78	·48	11
60—	53	·20	13
65—	33	·09	13·5
70—	15	·06	10·5
75—	6	·04	7
80—	2	·04	4

(14) What do you understand by a weighted average? Estimate the average number of children of married men of all ages from the data in the above table.

(15) Estimate the number of married men between the ages 52 and 53 in the same table, and also estimate at what age the average number of children is a maximum. Illustrate each estimate by a diagram.

(16) Define frequency group and standard deviation. Show that, if m'_2 is the second moment of any frequency group about any

origin and m_2 the second moment about the average of the group, and $\bar{x}$ is the average measured from the origin, then $m_2 = m'_2 - \bar{x}^2$. Calculate the standard deviation of the ages of *widowers*, shown in the table of question (13).

(17) State the product-sum formula for the correlation coefficient, and prove that if the means of rows and of columns of the correlation table lie on two straight lines the equations to these lines are

$$x = r\frac{\sigma_x}{\sigma_y}y, \qquad y = r\frac{\sigma_y}{\sigma_x}x,$$

respectively, x and y being deviations of the variables from their arithmetic means, r the correlation coefficient, and σ_x, σ_y the standard deviations.

(18) Below are given the populations of the County of London and the four surrounding administrative counties at the Censuses of 1891 and of 1901 :—

County.	Population.	
	1891.	1901.
London	4,228,317	4,536,541
Essex	578,471	816,640
Middlesex	542,894	792,314
Surrey	419,115	519,654
Kent	807,328	936,240
Total	6,576,125	7,601,389

(*a*) Assuming a constant percentage-rate of increase in each administrative county, estimate its population in 1896 at a date midway between the two Censuses.

(*b*) Assuming a constant percentage-rate of increase for the area as a whole (London and the four surrounding counties), estimate the total population at the same date in 1896.

Why does your estimate (*b*) differ from the sum of the estimates under (*a*) ?

(19) Give as exact a definition as possible of the term 'Cost of Living.' How far can the change in the Cost of Living be measured over a period in which there have been considerable modifications of diet or other changes in consumption of necessary commodities ?

(20) Discuss the methods of presenting wage statistics by averages. Illustrate by a diagram the following data :—

Building Trades.—Men, *Full time earnings.* Median, 37s. ; Quartiles, 29s. 6d., 40s. 6d. ; 5·9 per cent. received less than 20s. ; 2·8 per cent. received 45s. or more.

Estimate the average wage roughly.

(21) Construct a diagram to show graphically the relationship between yield of corn and rainfall from the data in the table below.

Years.	Yield per acre of corn, in bushels.	Rainfall in July, in inches.	Years.	Yield per acre of corn, in bushels.	Rainfall in July, in inches.
1886	24·5	1·15	1896	40·5	6·17
7	19·2	2·40	7	32·5	3·59
8	35·7	3·83	8	30·0	2·84
9	32·3	4·45	9	36·0	3·42
1890	26·2	2·03	1900	37·0	4·15
1	33·5	1·88	1	21·4	2·63
2	26·2	3·71	2	38·7	4·78
3	25·7	2·20	3	32·2	3·41
4	28·8	1·58	4	36·5	5·23
5	37·4	6·01	5	39·8	4·78

(22) *Find the correlation between yield of corn and rainfall in the above table.*

(23) Define (*a*) arithmetic average, (*b*) geometric average, (*c*) median, (*d*) mode, (*e*) quartile.

Instance cases when (*b*), (*c*) and (*d*) are specially appropriate.

(24) Comment on the form of grouping adopted in (26) Table I., and state any inconveniences that it presents.

Calculate approximately the values of the median and quartiles, using a graphic method.

(25) Explain what is meant by the skewness of a frequency distribution. Give Pearson's measure of the skewness, and any other way of measuring it.

Obtain some measure of the skewness of the distribution in (26) Table II.

(26) Table I., showing the number of civil parishes in England and Wales in which the population at the Census of 1901 lay between the limits given in the column on the left :—

Population.	Number of Civil Parishes.	Population.	Number of Civil Parishes.
None	25	500 and under 750	1,557
1 and under 50	812	750 „ 1,000	842
50 „ 100	1,339	1,000 „ 5,000	2,411
100 „ 200	2,503	5,000 „ 10,000	413
200 „ 300	2,036	10,000 „ 20,000	241
300 „ 400	1,410	20,000 and upwards	273
400 „ 500	1,038		
		Total No. of Civil Parishes	14,900

Table II., showing the number of rooms measured, in a certain investigation, in which the size lay between the limits given in the column on the left; area calculated to the nearest square foot :—

Area of Room in Square Feet.	Number.	Area of Room in Square Feet.	Number.
20 and under 40	3	200 and under 220	18
40 ,, 60	14	220 ,, 240	12
60 ,, 80	16	240 ,, 260	3
80 ,, 100	36	260 ,, 280	2
100 ,, 120	31	280 ,, 300	2
120 ,, 140	35	300 ,, 320	2
140 ,, 160	35	320 ,, 340	..
160 ,, 180	15	340 ,, 360	..
180 ,, 200	26	360 ,, 380	1
		Total No. of Rooms Measured	251

Write a short account of the use of graphic methods in statistics. Draw diagrams representing the data of Table I. and of Table II.

(27) Define the standard deviation, and show that the mean square deviation is least when deviations are measured from the arithmetic mean.

Find the mean and standard deviation for the sizes of the rooms given in (26) Table II.

(28) What corrections are applied to the crude death-rates of areas in order to obtain comparable rates ?

(29) (1) Estimated average weekly wages of agricultural labourers in thirty-six counties of England in 1891, and (2) the percentage of the population in receipt of poor law relief, in rural unions of the same counties, on 1st January of the same year :—

County.	Wages. s. d.	Percentage in Receipt of Relief.	County.	Wages. s. d	Percentage in Receipt of Relief	County.	Wages. s. d.	Percentage in Receipt of Relief.
1	18 6	1·7	13	14 8	3·6	25	12 0	4·9
2	18 0	2·3	14	14 0	3·1	26	12 0	4·7
3	17 0	2·5	15	14 0	4·0	27	12 0	3·9
4	17 0	2·1	16	14 0	2·3	28	11 6	4·0
5	16 0	3·0	17	13 0	2·8	29	11 6	4·5
6	16 0	2·1	18	12 0	4·5	30	11 6	4·2
7	15 6	2·8	19	12 0	4·7	31	11 0	5·2
8	15 6	2·7	20	12 0	4·7	32	11 0	4·2
9	15 0	3·5	21	12 0	5·7	33	10 6	4·2
10	15 0	3·1	22	12 0	4·1	34	10 0	3·2
11	15 0	3·1	23	12 0	4·9	35	10 0	4·4
12	15 0	2·7	24	12 0	4·2	36	10 0	4·8

Define the arithmetic mean, the median and the mode, and give a sketch of a skew frequency distribution showing the approximate position of each.

State the chief advantages of the arithmetic mean as a form of average, and find the arithmetic mean and the median for the wages of agricultural labourers in the above table.

(30) Explain clearly the meaning of the term 'dispersion,' and find the mean deviation from the median for wages in the same table.

(31) Also, define standard deviation, and find the standard deviation of these wages.

(32) Using the data in question (29), test graphically, with squared paper, the correlation between average wages and percentages of the population in receipt of poor law relief, stating your conclusions in words.

(33) Construct a blank table, complete with headings and lines, and with due regard to spacing, in which could be inserted the numbers of persons employed in six groups of industries, four grades of age at three different periods.

(34) The following table gives for 780 weeks the call discount rate and the ratio of reserves to deposits in New York. Calculate the average discount rate for the various ratios of reserves to deposits, and express the results graphically.

	Call Discount Rates.														
Ratio of Reserves to Deposits.	1-	2-	3-	4-	5-	6-	7-	8-	9-	10	12-	15-	20-	25-	Totals.
21°/₀-	...	...	...	1	1	...	...	...	1	...	...	...	...	...	3
23°/₀-	...	...	...	1	1	2	...	3	...	2	...	...	...	1	10
25°/₀-	...	...	9	33	30	27	9	6	3	4	1	2	1	2	127
27°/₀-	6	72	57	42	36	16	7	1	...	1	...	1	...	...	239
29°/₀-	25	87	31	13	3	...	2	...	...	...	...	...	...	1	162
31°/₀-	47	26	11	4	1	...	...	...	...	...	...	...	...	...	89
33°/₀-	36	6	1	3	...	...	...	...	...	...	...	...	...	...	46
35°/₀-	22	2	...	...	...	...	...	...	...	...	...	...	...	...	24
37°/₀-	18	2	...	...	...	...	...	...	...	...	...	...	...	...	20
39°/₀-	36	...	...	...	...	...	...	...	...	...	...	...	...	...	36
41°/₀-	20	...	...	...	...	...	...	...	...	...	...	...	...	...	20
43°/₀-	2	...	...	...	...	...	...	...	...	...	...	...	...	...	2
45°/₀-	2	...	...	...	...	...	...	...	...	...	...	...	...	...	2
Totals, .	214	195	109	97	72	45	18	10	4	7	1	3	1	4	780

The heading 15- covers all rates of 15 and over but less than 16, etc.

From the following data find the equation of the regression line giving the average discount rate for all ratios of reserves to deposits,

and plot the line on the same diagram. Is the use of the product moment method of determining correlation justifiable in this case ?

	Means.	Standard Deviations.	Correlation.
Call Discount Rates . .	3·6	2·5	−·52
Ratios of Reserves to Deposits	30·3	4·2	

(35) As an illustration of the nature of definitions in statistics explain fully the meaning of the statement: 'The total value of exports (produce and manufactures of the United Kingdom) in 1918 was £498,473,065.'

PART II

(1) Give a short account of the chief official publications, in England, relating to statistics of *one* of the following subjects, with especial reference to the source and the precise meaning of the data :—

(*a*) Vital statistics (births, deaths and marriages).
(*b*) Foreign trade.
(*c*) Agriculture.

(2) What do you understand by the words 'frequency group' ?

England and Wales, 1911

Ages	10–	15–	20–	25–	35–	45–	55–	65–
All Occupied (Males 000s.)	246	1164	1146	2225	1815	1262	723	299
Coal Hewers (00s.)	63	338	538	1067	798	467	212	50

Compute suitable averages and measures of dispersion for the comparison of the age groups in this table and comment on the results.

(3) What means are available for testing the significance of differences between statistical coefficients ? Test whether the differences between the means and measures of dispersion for the two series given in the previous question are significant.

(4) $s_1=4{\cdot}53$ and $s_2=3{\cdot}71$ are the standard deviations of two groups, $x_1, x_2 \ldots x_n$, and $y_1, y_2 \ldots y_n$. $Sxy=4832$. $n=1000$.
Explain exactly the meaning of standard deviations. Calculate the product-sum coefficient of correlation between the groups, and state what it measures. Write down the probable error of the coefficient and explain its meaning.

(5) Find the standard deviation of the differences between corresponding values of two variates x and y.

(6) Set out in detail the method by which you would make graphic comparisons of two such series of figures as Imports of Manufactures and Unemployment.

(7) If the recorded births in a certain district may be in defect by x per cent., and the estimated population in error by $\pm y$ per cent., find an approximate expression for the greatest possible error in the birth-rate, x and y being assumed fairly small (say, not more than 5 per cent. or so).

(8) Given five thousand different figures—*e.g.* quotations of prices, or measurements of human statures—how would you (*a*) select five hundred figures at random from that total, and (*b*) ascertain the probability that the average of the five hundred selected figures does not differ from the average of the five thousand by more than any assigned extent ?

(9) (*a*)

Number of Persons per Tenement.	Number of Rooms per Tenement.									Total.	Approx. Average Number of Rooms.
	2	3	4	5	6	7	8	9	10 or more.		
1	26	9	8	5	1	4	1	1	1	56	3·5
2	14	24	60	40	29	9	8	3	4	191	4·8
3	5	24	61	57	34	11	8	3	5	208	5·0
4	4	16	57	44	27	16	3	2	10	179	5·2
5	3	18	36	42	21	6	5	5	12	148	5·3
6	2	8	21	43	19	4	5	3	7	112	5·4
7	...	5	18	14	11	5	3	2	4	62	5·5
8	...	1	14	11	10	...	2	2	4	44	5·7
Totals,	54	105	275	256	152	55	35	21	47	1000	5·1
Average Number of Persons,	2·07	3·56	3·94	4·24	4·24	3·78	4·14	4·62	4·81	3·976	

Standard deviations: persons 1·83, rooms 1·9 (approx.).

(*b*) Show that the coefficient of correlation can be expressed in the form

$$\frac{1}{\sigma_1\sigma_2}\left\{\frac{1}{n}S(xy)-\bar{x}\bar{y}\right\},$$

where $\bar{x}$, $\bar{y}$ are the averages of the observations referred to any origin.

(*c*) Calculate, by any method, a measurement of the correlation between the number of rooms and number of persons per tenement shown in (*a*).

(*d*) Calculate the third and fourth moments of the frequency curve of persons; determine the position of the mode and also determine the skewness by any method known to you.

(10) The table given below gives the results of the measurement of series of 959 Oxford Students and of 2348 convicts. Find what, if any, differences between the statistical constants given are significant and comment on the results.

Character.	Data.	Means.	Standard Deviations.	Coefficients of Correlation with Stature.
Head-length	Students	196·05	6·23	·31
	Convicts	192·44	6·39	·26
Head-breadth	Students	152·84	4·92	·14
	Convicts	151·02	5·49	·15
Head-height	Students	136·62	5·80	·28
	Convicts	132·29	5·21	·19
Stature	Students	69·49	2·60	—
	Convicts	65·44	2·65	—

(11) Give a short account of the nature of the information contained in *one* of the following : Census of Production, 1907 ; Reports on Wages in 1906 (the 'Wage Census') ; Reports on Buildings and Tenements (housing and overcrowding) in the Population Census, 1911.

(12) Outline a method by which the normal curve of error can be obtained as the limit of $(p+q)^n$.

(13) Discuss (*a*) the best means of obtaining accurate statistics of family expenditure, and (*b*) the best means of combining such data so as to form a representative type.

(14) Define the following terms and give illustrations of their use : interpolation, standard deviation, moment, skewness, logarithmic scale, geometric mean, partial correlation, normal curve of error.

(15) In m trials an event has happened r times. How would you determine the probability that this result is consistent with the hypothesis of *random* sampling from a universe in which the chance of the event happening is a certain *small* quantity p ? Why cannot the required probability be derived from a table of the normal curve of error ?

(16) If m_1, m_2 are the numbers of deaths occurring in a year among N_1, N_2 persons of two different occupations, the standard deviation by which the significance of the observed difference in the death rates per 1000 can be tested is given as

$$1000\sqrt{\left\{\frac{m_1(N_1-m_1)}{N_1^3}+\frac{m_2(N_2-m_2)}{N_2^3}\right\}}.$$

Show how this formula is obtained and criticise it.

(17) Contrast the methods used in the construction of any two current index numbers of wholesale prices. Under what conditions is 'weighting' important in index numbers ?

(18) Analyse in some detail the cases in which it may be assumed (1) that a frequency distribution is normal, or (2) that the probability of errors in a measurement or observation exceeding various amounts is determinable by the normal table of probability.

(19) What methods are available for testing the 'goodness of fit' of a mathematical curve to observations ?

(20) If $z=x_1+x_2+ \ldots +x_n$, where the x's are deviations from the average of quantities selected at random and independently of each other from a curve frequency whose standard deviation is σ, show that the standard deviation of z is $\frac{\sigma}{\sqrt{n}}$, n being finite.

Under what circumstances can it be shown that the curve of frequency of z is normal ?

(21) What methods are available for classifying frequency curves into types ? State briefly the mathematical concepts underlying the Pearsonian classification of frequency curves.

(22) Explain how the necessity for Sheppard's corrections of moments arises.

If m_2 is the second moment calculated from observations when all are supposed to be grouped at the middle of grades whose breadth is h, show that $m_2+\frac{h^2}{12}$ is the second moment if it is assumed that the observations are evenly distributed through the grades.

(23) A sample containing 1000 is drawn at random from a large universe and 300 are found to possess a certain attribute. Can you infer anything as to the proportion in the universe that have this attribute, or what further information is needed?

(24) Discuss the effect on a weighted mean of errors in the quantities or the weights.

(25) Write a brief note on the assumptions made in calculating the probable error of a statistical quantity, such as the standard deviation or the correlation coefficient.

(26) Calculate the average, second, third and fourth moments, mean deviation, standard deviation and skewness of the frequency groups of chest girths shown in the following table :—

Height and Chest Girth of 1126 Recruits of 18 Years of Age.

Chest Girth in Inches.	Height in Inches.													Totals.
	60	61	62	63	64	65	66	67	68	69	70	71	72	
28	...	...	1	...	...	...	...	...	...	...	...	...	...	1
29	...	...	1	...	...	1	...	...	...	...	...	...	...	2
30	1	1	3	...	1	...	1	...	2	...	...	...	...	9
31	2	9	8	3	4	6	7	3	2	2	...	...	1	47
32	8	18	24	29	36	12	16	5	8	2	...	1	...	159
33	6	11	21	30	42	22	36	17	8	9	1	3	1	207
34	6	16	15	43	52	43	21	40	28	12	1	3	...	280
35	...	6	15	25	32	32	29	29	18	16	14	2	1	219
36	...	4	3	6	11	22	18	18	18	19	6	2	...	127
37	...	1	...	3	1	4	12	6	8	6	6	1	1	49
38	...	2	...	1	1	1	4	3	4	4	...	2	...	22
39	...	...	...	...	...	...	...	...	1	1	2	...	...	4
Totals, . .	23	68	91	140	180	143	144	121	97	71	30	14	4	1126
Averages of Arrays, .	32·7	33·6	33·5	34·2	34·1	34·7	34·7	35·0	35·1	35·5	36·3	35·4	34·5	34·51
Standard Deviations of Arrays, .	1·75	1·87	1·57	1·34	1·33	1·50	1·76	1·40	1·77	1·67	1·3	1·8	2·2	1·66

Average height, 65·6 ; standard deviation, 2·52.

Compare the relations between the quantities calculated with those that are found in the normal curve of error.

(27) From the above data draw the regression line (chest girth on height), and with the help of your drawing find an approximate value of the coefficient of correlation between height and chest girth.

In normal distributions, the standard deviation of an array is $\sigma_2\sqrt{1-r^2}$ where σ_2 is the standard deviation of the arrays merged in one group. Are the standard deviations shown in the table consistent with this formula?

INDEX

[*The numbers refer to pages ; I and II refer to the two parts of the book.*]

ABSCISSA, I 68.
Abstract of Labour Statistics, I 19 ; II 177, 213, 283.
Advancing Differences, I 92.
Age Distribution of Criminals, I 94.
Anemone Nemorosa, II 227-30.
Arithmetic Mean (*see* Mean).
Array, I 109, 122-3, 127 ; II 251, 256-7, 262.
Assortative Mating, II 168.
Average, I 4, 9-11, 22-41, 63, 115, 125 ; II 136, 155-7, 173.

BERNOULLI, II 149, 248.
Beveridge, I 15, 81.
Biassed and Unbiassed Errors, II 154-6, 170, 179.
Binomial, I 91 ; II 141-3, 146-7, 151, 181, 187-8, 190, 232-5, 240.
Biometrika, I 51, 103, 109, 114 ; II 134, 161-3, 165-7, 196, 212, 227, 230.
Birth Rate, I 7, 32, 82 ; II 234.
Board of Trade Index Number of Prices, I 39.
Board of Trade *Journal and Commercial Gazette*, II 282.
Booth, II 279.
Bowley, I 37 ; II 170-3, 280.
Boyle's Law, I 75.
Brain-weight, I 103-4 ; II 166.
Bravais, I 114.
Burnett-Hurst, II 173.

CALCULATION of Mean and Standard Deviation, I 52-5.
Cattle and Grass-land, I 120-5.
Census, I 1, 14-15, 20, 40, 85, 117 ; II 177, 280-2.
Central Statistical Office, II 279.
Centre of Gravity, I 60 ; II 269.
Chance, I 4 ; II 134, 136.
Charlier, II 248.
Classification and Tabulation, I 4, 14-21.
Class-interval, I 21, 52, 103, 120-3 ; II 195, 213, 218.
Coefficient of Correlation, I 108, 110-12, 120, 129, 131 ; II 133, 152, 158, 162-3, 168-9, 253, 260, 278.
Coefficient of Variation, I 50-51 ; II 133-4, 162-4, 167-8.
Compound Interest, I 7 ; II 263.
Condition of School Children, I 17.
Constancy of Great Numbers, I 1.
Consumers' Surplus, I 97, 101.
Continuous Variation, I 26 ; II 182, 195-6, 232.
Co-ordinates, I 68.
Corrected Death Rate, I 32-5.
Correlation, I 4, 18, 76, 78, 82, 102-31 ; II 151-3, 166-7, 253.
Correlation Ratio, I 112, 114 ; II 163.
Cost of Living, I 8, 29, 35-8, 125-31 ; II 281-2.
Criminal Anthropometry, II 166.
Crude Death Rate, I 32-5.
Cuckoo's Eggs, II 161.
Cunynghame, I 95.
Curve Fitting, II 179-230.

DARWIN, I 3.
Death Rate, I 7, 32-5, 82 ; II 234.
Deciles, I 49.
Decreasing Return, I 99-100.
De Moivre, II 149.
Descartes, I 68.
Dispersion, I 4, 42-51, 108 ; II 145, 162.
Distribution of Random Digits, II 157.
Dividend Sample, II 171.

Economist, I 9-11.
Edgeworth, I 3, 114 ; II 172, 230 248.
Elasticity of Wire, I 73.
Elderton, II 190, 203, 230.
Ellipse, I 69 ; II 252, 257-62.
Error, I 105.
Errors of Observation, I 4, 73 ; II 233.
Euler, II 149, 248.
Examination Marks, I 25, 27-8, 52-5, 66-7, 93 ; II 175-6, 207-212.

Fechner, II 248.
Fermat, II 149.
Fitting of Curves, II 179-230.
Fitting with a Parabola, I 88-9.
Fluctuations of Sampling, I 47.
Flux, I 39.
Frequency, I 12, 54, 60, 62, 102-3, 122-3, 127; II 135, 143, 150-3, 157, 172, 178-9, 184, 195-6, 200, 238-9.
Frequency Curve, I 60; II 178-230.
Frequency Distribution, I 11-13, 52-67; II 194, 267.
Frequency Polygon, II 180, 182, 184, 195.
Frequency Surface, II 249-62.
Function, I 72, 87.

Galton, I 3, 49, 108-9, 114; II 248.
Gamma Function, II 214.
Gauss, II 248, 273.
Generalised Probability Curve, II 187.
Geometric Mean, I 38-9; II 264-5.
Goodness of Fit, II 180, 212, 219, 223-224.
Graphs, I 57, 68-101.

Hain, I 3.
Halley, I 1.
Harmonic Mean, I 39; II 264-5.
Height of School Children, I 18.
Herschel, II 234.
Histogram, I 60; II 180, 219-20, 223, 225, 227, 232.
Homogeneity, II 165-6.
Household Budgets, I 35-6; II 155, 266.
Hyperbola, I 69.
Hypergeometrical Expansion, II 187.

Increasing Return, I 100.
Index Numbers, I 8-11, 29-31, 35-9, 76-80; II 156.
Indictable Offences and Unemployment, I 82-3.
Infant Mortality, I 65, 117-19.
Infectious Disease Rate, I 56, 62; II 176, 219-27.
Inheritance, II 162, 167-8.
International Statistical Congress, I 2.
Interpolation, I 24, 48-9, 76, 85-94; II 212.

Journal of the Royal Statistical Society, I 8, 37; II 170, 230.

Kapteyn, II 248.
Knapp, I 3.

Labour Gazette, I 120; II 282.
Lagrange, II 149.
Lagrange's Interpolation Formula, I 94.
Laplace, II 149, 248.
Latter, II 161.
Least Squares, I 105-6; II 196, 271-3.
Lee, I 109; II 167.
Levis, I 3.
L'homme Moyen, I 2.
Limits for Correlation Coefficient, I 110.
Lipps, II 248.
Livelihood and Poverty, II 173.
Logarithmic Curve, I 87.
London Statistics, I 83, 117.

McAlister, II 248.
Macdonell, I 51; II 166.
Maclaurin, I 87; II 187.
Marriage Rate, I 32, 77-81, 115-17; II 234.
Marshall, I 95.
Mean—Arithmetic Mean, I 22, 30, 38-41, 52-5, 60, 62-5, 104-13, 116-19, 122-5, 127-9; II 132-4, 144-5, 147-8, 153, 157-8, 160-4, 167-9, 193, 201, 204-5, 208, 217, 223, 226, 228, 238, 251, 256-8, 265-9, 273.
Mean Deviation, I 42-6, 50, 55, 64; II 246, 270-1.
Mean Error, II 245.
Mean-square Deviation, I 46-7, 54; II 211, 268-9.
Median, I 23-6, 36, 38-41, 43-6, 60-7, 87; II 162, 204, 238, 270.
Median Error, II 245.
Meteorological Observations, I 83-4.
Minority Report on the Poor Law, I 15.
Mode, I 26-9, 36, 38-41, 60-2, 64, 66-7; II 162, 186, 190-3, 204-5, 217, 223, 226, 228-9, 238.
Moments, I 123-4, 127; II 152-3, 163, 187, 194-205, 207-8, 213-16, 220-22, 225-6, 228, 267-9.
Monopoly, I 101.
Monthly Bulletin of Statistics, II 283.

Non-linear Regression, I 112, 114.
Non-measurable Characters, I 6; II 162.
Normal Curve of Error, I 4; II 133, 184, 193, 206-12, 231-48, 251-2, 256.
Normal Demand, I 99.

Occupational Death Rate, I 34.
Ogive Curve, I 66.
Ordinate, I 68.
Overcrowding, I 20, 117-19.

Paish, I 9.
Parabola, I 69, 73, 88-9; II 188, 197-9.
Pascal, II 149.
Pearl, I 103; II 166.
Pearl and Dunbar, I 51.
Pearl and Fuller, II 134.

Pearson, I 3, 50-1, 60-1, 108-9, 112, 114; II 162, 165, 167, 184, 187, 194, 212, 230, 248, 255, 276.
Petty, I 1.
Philosophical Transactions, II 162.
Plotting of a Frequency Distribution, I 55-60.
Point of Inflexion, II 238, 245.
Poisson, II 248.
Poor Relief, I 88-92, 120.
Population according to Age, I 20.
Prediction, II 169.
Prices, I 8-11, 35-9, 76-81, 115-17, 125-31; II 235, 265, 281-2.
Probability, I 3-4; II 132-49, 151, 181, 187, 231, 236-50.
Probability Curve, I 4; II 184, 187, 236.
Probability Integral, II 209, 239, 245, 275.
Probable Error, II 133-4, 145, 150-64, 245-6.
Probable Error of Mean, II 153-4.
Probable Error of Sum or Difference, II 157-8.
Proceedings of London Mathematical Society, II 203.
Producers' Surplus, I 98, 100-1.
Product Deviation, I 106, 113, 116, 118-9, 123-5, 127-9.
Production Census, II 177, 282.
Progress of the Nation, I 88, 94.

Quadrature Formulae, II 197, 209, 218, 230.
Quartile, I 48-9, 64, 66-7, 87; II 245.
Quartile Deviation, I 48-50, 55, 64-5; II 245-6.
Questionnaire, I 15.
Quetelet, I 1-3.

Random Errors, I 4.
Random Sampling (*see* Sampling).
Registrar-General, I 3, 14, 33, 55; II 280, 283.
Regression, Coefficient of, Line of, I 108-12, 117, 119, 125-6, 129-31; II 162, 253, 257, 260-1.
Rent, I 29-31, 101, 125-31; II 281.
Rowntree, II 173, 279.
Royal Society Transactions, II 230.

Sampling, Random Sampling, I 4, 47; II 132-4, 145-77, 179, 182, 212, 236, 243.
Sauerbeck, I 8-10, 35.
Schuster, II 163.
Secular Trend, I 80.
Sheppard's Adjustment, I 104; II 203, 208, 213, 225.
Short-time Fluctuations, I 78, 80.
Significance, II 133, 159-68.
Simpson's Rule, II 219, 230.
Skew, Skewness, I 4, 49, 61-4; II 146, 180, 184, 186-7, 190-2, 205, 235, 247.
Sleep and Physical Condition, I 19.
Smooth Curve, I 60, 86; II 234.
Social Statistics, II 279.
Standard Deviation, I 46-7, 50-5, 63, 106-13, 116-19, 122-5, 127-9; II 133-4, 145-8, 151-4, 157-64; 166-8, 171-2, 176-7, 207-9, 240-6, 251, 256-8, 266-9.
Standard Error, II 134.
Standard Population, I 34.
Statical Analogues, II 269, 271.
Statist, I 9.
Stirling, II 248.
Successful Events, II 136-8, 140-8, 151-4, 169, 171, 174-5, 231, 240.
Successive Differences, I 91.
Supply and Demand Curves, I 95-101.
Süssmilch, I 1.
Symmetrical Distribution, I 61; II 180.

Table of Probable Errors, II 163.
Tawney, I 40; II 158.
Tax, I 100-101.
Temperature Records, I 65.
Todhunter, II 149, 248.

Unemployment, I 15, 81-3; II 213-19.

Variability, I 42-51, 61, 63, 107-108, 111; II 132-4, 162, 165-6, 247-8.
Variable, I 6-13, 72, 75, 87, 116, 122; II 263-9.
Variation, I 6.
Variation in Arcella, I 51.
Variation in the Earthworm, II 134.
Variation in Eupagurus Prideauxi, II 163.
Variation in Plants, II 167.

Wages, I 8, 40-2, 125-31; II 173, 224, 227, 234, 281-3.
Weighting, I 30-8.
Weighted Mean, I 29-38, 263-4.

Yule, I 114; II 227.

LIBRARY
1380
AGRICULTURAL COLLEGE

Uniform with this Volume

BELL'S MATHEMATICAL SERIES

(Advanced Section)

General Editor: **WILLIAM P. MILNE, M.A., D.Sc.**
Professor of Mathematics, Leeds University

AN ELEMENTARY TREATISE ON DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS

By H. T. H. PIAGGIO, M.A., D.Sc., Professor of Mathematics, University College, Nottingham; formerly Senior Scholar of St. John's College, Cambridge. Third Edition. Demy 8vo. 12*s*. net.

The earlier chapters contain a simple account of ordinary and partial differential equations, while the later chapters are of a more advanced character, and cover the course for the Cambridge Mathematical Tripos, Part II., Schedule A, and the London B.Sc. Honours. The number of examples, both worked and unworked, is very large, and all the answers are given.

'With a skill as admirable as it is rare, the author has appreciated in every part of the work the attainments and needs of the students for whom he writes, and the result is one of the best mathematical text-books in the language.'—*Mathematical Gazette.*

A FIRST COURSE IN NOMOGRAPHY

By S. BRODETSKY, M.A., B.Sc., Ph.D., Professor of Applied Mathematics at Leeds University. Demy 8vo. 10*s*. net.

Graphical methods of calculation are becoming of ever greater importance in theoretical and industrial science, as well as in all branches of engineering practice. Nomography is one of the most powerful of such methods, and the object of this book is to explain what nomograms are, and how they can be constructed and used. The book caters for both the practical man who wishes to learn the art of making and using nomograms, and the student who desires to understand the underlying principles. It is illustrated by sixty-four figures, most of which are actual nomograms, their construction being analysed in the text. In addition, there are numerous exercises illustrative of the principles and methods.

'A good introductory treatise . . . calculated to appeal to the student who desires to make early practical use of the knowledge he acquires.'
The Mechanical World.

ELEMENTARY VECTOR ANALYSIS

With application to Geometry and Physics. By C. E. WEATHERBURN, M.A., D.Sc., Professor of Mathematics, Canterbury University College, Christchurch, N.Z. Demy 8vo. 12*s*. net.

This book provides a simple exposition of Elementary Vector Analysis, and shows how it may be employed with advantage in Geometry and Mathematics. The use of Vector Analysis in the former is abundantly illustrated by the treatment of the straight line, the plane, the sphere and the twisted curve, which are dealt with as fully as in most elementary books. In Mechanics the author has explained and proved all the important elementary principles.

LONDON: G. BELL AND SONS, LTD.

BELL'S MATHEMATICAL SERIES—*Contd.*

(Advanced Section)

ADVANCED VECTOR ANALYSIS

With Application to Mathematical Physics. By C. E. WEATHERBURN, D.Sc., Professor of Mathematics, Canterbury University College, Christchurch, N.Z. Demy 8vo. 15*s.* net.

'The author has already published in this series an *Elementary Vector Analysis.* In this companion volume he deals with the higher part of the subject and the applications of the theory, including a chapter on the Equations of Maxwell and Lorentz and the Lorentz-Einstein transformation.'—*Times.*

THE ELEMENTS OF NON-EUCLIDEAN GEOMETRY

By D. M. Y. SOMMERVILLE, M.A., D.Sc., Professor of Mathematics Victoria University College, Wellington, N.Z. 7*s.* 6*d.* net.

'An excellent text-book for all students of Geometry.'—*Nature.*
'A useful and stimulating book.'—*Mathematical Gazette.*

ANALYTICAL CONICS

By D. M. Y. SOMMERVILLE, M.A., D.Sc. Demy 8vo. 15*s.* net.

In this book the elementary properties of the Conics are first of all dealt with, and thereafter the higher portions of the subject, such as Conics referred to any axes, Homogeneous Co-ordinates, Invariant and Covariant properties of Conics, etc. There are abundant collections of examples in all the subjects treated.

A Professor of Mathematics writes: 'I find it a work entirely praiseworthy; and indeed of such excellency as one would expect from the pen of the author of *The Elements of Non-Euclidean Geometry.* The variety of topics treated is more extensive than that to be found in most existing text-books on the subject. In certain instances the treatment is refreshingly novel, and in all cases the presentation is concise and lucid.'

A TEXT-BOOK OF GEOMETRICAL OPTICS

By A. S. RAMSEY, M.A., *President of Magdalene College, Cambridge.* New and Revised Edition. Demy 8vo. 8*s.* 6*d.*

This is a revised edition of the work published in 1914. It contains chapters on Reflection and Refraction, Thin and Thick Lenses, and Combinations of Lenses, Dispersion and Achromatism, Illumination, The Eye and Vision, Optical Instruments, and a chapter of Miscellaneous Theorems, together with upwards of three hundred examples taken from University and College Examination Papers. The range covered is somewhat wider than that required for Part I. of the Mathematical Tripos.

LONDON: G. BELL AND SONS, LTD.
YORK HOUSE, PORTUGAL ST., W.C. 2.